electronics

MDPI

Article

Trust by Design: An Ethical Framework for Collaborative Intelligence Systems in Industry 5.0

Emmanuel A. Merchán-Cruz [1,*], Ioseb Gabelaia [2], Mihails Savrasovs [1], Mark F. Hansen [3], Shwe Soe [3], Ricardo G. Rodriguez-Cañizo [4] and Gerardo Aragón-Camarasa [5]

[1] Engineering Faculty, Transport and Telecommunication Institute, Lauvas Iela 2, LV-1019 Riga, Latvia; savrasovs.m@tsi.lv
[2] School of Business, Graceland University, Lamoni Campus, 1 University Pl, Lamoni, IA 50140, USA; gabelaia@graceland.edu
[3] Bristol Robotics Laboratory, University of the West of England, Bristol BS16 1QY, UK; mark.hansen@uwe.ac.uk (M.F.H.); shwe.soe@uwe.ac.uk (S.S.)
[4] Escuela Superior de Ingeniería Mecánica y Eléctrica—Unidad Azcapotzalco, Instituto Politécnico Nacional, Av. de las Granjas 682, Mexico City C.P. 02550, Mexico; rgrodriguez@ipn.mx
[5] School of Computing Science, University of Glasgow, 18 Lilybank Gardens, Glasgow G12 8RZ, UK; gerardo.aragoncamarasa@glasgow.ac.uk
* Correspondence: merchan.e@tsi.lv

Abstract: Industry 5.0 highlights human-centricity, sustainability, and resilience. This article presents a novel Trust by Design framework applicable to collaborative intelligence systems within Industry 5.0, addressing the need for collaborative systems to be reliable by design, incorporating ethical principles such as transparency, accountability, fairness, and privacy throughout the entire system lifecycle. The framework is grounded in select ethical philosophies applied to practical design requirements for human-AI collaboration, identifying key ethical challenges that threaten to damage trust and restrict the adoption of collaborative systems. The authors employ a qualitative, literature-driven method, conceptual modeling, and scenario-based case study analysis, synthesizing best practices and ethical policies from the EU AI Act, GDPR, and more. Trust by Design suggests a structured set of principles and implementation measures to embed ethics into every phase of the system's lifecycle. The applicability and suitability of the framework are demonstrated through representative real-world application scenarios across industries. The results indicate that trust in collaborative intelligence systems is not static but dynamic, context-dependent, and controlled by transparency, fairness, and user experience. The framework includes instruments and methods to measure ethical performance, including trust metrics, override rates, fairness indicators, and incident tracking.

Keywords: human-centric AI; AI governance; human–robot collaboration; ethical framework; responsible AI

Academic Editor: Chang Wook Ahn

Received: 15 April 2025
Revised: 30 April 2025
Accepted: 8 May 2025
Published: 11 May 2025

Citation: Merchán-Cruz, E.A.; Gabelaia, I.; Savrasovs, M.; Hansen, M.F.; Soe, S.; Rodriguez-Cañizo, R.G.; Aragón-Camarasa, G. Trust by Design: An Ethical Framework for Collaborative Intelligence Systems in Industry 5.0. *Electronics* **2025**, *14*, 1952. https://doi.org/10.3390/electronics14101952

1. Introduction

With the rapid pace that artificial intelligence is not just being developed but also adopted into all kinds of systems, Industry 5.0 represents the evolution of industrial transformation that emphasizes the close interaction between humans and advanced technologies beyond the automation focus of Industry 4.0 [1,2]. Under this paradigm, the concept of collaborative intelligence is central to achieving enhanced productivity and innovation that involves humans and AI systems working together complementarily so that each augments the other's strengths [3–6]. However, to achieve the necessary level

of synergy between human collaborators and intelligent systems, a number of ethical considerations need to be addressed so that trust among the different actors becomes natural; otherwise, issues such as adoption resistance, underuse of the technology, or even misuse and accidents can hinder further integration [7]. Consequently, there is a clear need to define an ethical framework that guides the design and deployment of these systems so that they can be trustworthy by design, ensuring a correct alignment with human values, rights, and expectations from their very inception.

The need for ethical frameworks in human-machine collaboration is driven by concerns about transparency, privacy, autonomy, and accountability in increasingly complex socio-technical scenarios [8,9]. Without clear ethical guidelines, collaborative robots and AI decision aids may unintentionally undermine user autonomy or introduce biases, contradicting the human-centric goals of Industry 5.0. This paper introduces a comprehensive "Trust by Design" framework for collaborative intelligence systems, embedding ethical principles and trust-building mechanisms throughout the entire system lifecycle. By doing so, it aims to ensure that human-machine collaboration in Industry 5.0 remains both responsible and centered on human values.

We ground our framework in existing literature and ethical theories, adopting an interdisciplinary methodology. We begin by reviewing the emergence of Industry 5.0 and its core pillars, highlighting the gap in ethical governance. We then identify key ethical challenges in collaborative intelligence by synthesizing insights from technology ethics guidelines and stakeholder perspectives. Building on these insights, we propose the Trust by Design framework with clearly defined principles and actionable measures. The framework's implementation is further detailed through integration into development processes, risk assessment methods, and governance structures. To illustrate practical relevance, we examine general case studies across manufacturing, decision support, and human augmentation scenarios. Our research questions include:

(1) What are the unique ethical challenges of human-centric collaborative intelligence in Industry 5.0?
(2) How can systems be designed to inherently foster trust between humans and machines? and
(3) What governance and validation mechanisms ensure ongoing ethical compliance?

Considering that the goal of this study is to develop a conceptual framework, rather than a quantitative evidence aggregation, we employed a qualitative, literature-driven method based on a streamlined systematic review. This approach allows for the integration of diverse knowledge sources essential for comprehensive framework development.

A comprehensive search strategy was developed using multiple search strings across major scientific databases. The search was conducted between October 2024 and March 2025 on Scopus and IEEE Xplore, with the following search strings:

1. ("Industry 5.0" OR "Operator 5.0" OR "Society 5.0") AND ("human-robot collaboration" OR "human-machine collaboration" OR "human-AI collaboration" OR "collaborative intelligence" OR "human-centred" OR "human-centric")
2. ("trust" OR "ethics" OR "sustainability" OR "resilience") AND ("artificial intelligence" OR "robotics" OR "automation") AND ("human factors" OR "collaboration" OR "teaming")
3. ("value sensitive design" OR "ethics by design" OR "privacy by design") AND ("technology" OR "artificial intelligence" OR "industry")
4. ("collaborative robots") AND ("acceptance" OR "trust" OR "implementation")

All searches were limited to English-language articles and conference papers published between 2000 and 2025. Considering that concern for the introduction of robots has been

studied since the early 2000s, foundational works were identified while more recent sources were prioritized for relevance.

Initial database searches yielded 9211 records (Scopus: 7162; IEEE Xplore: 2049). After removing 1607 duplicate records (approximately 18%), 7604 unique records remained for screening. Based on predefined criteria that considered insufficient focus on Industry 5.0 or human-machine collaboration, purely technical content without human factors consideration, limited relevance to trust, ethics, or sustainability, and lack of collaborative scenarios, after screening the title and abstract, 7215 records were excluded.

Full-text assessment of the remaining 389 articles resulted in the exclusion of 296 articles due to insufficient focus on human-machine collaboration, limited discussion of trust or ethics, methodological limitations, and duplicate research or insufficient novelty.

To ensure comprehensive coverage, non-database records were added, including policy documents and frameworks, industry white papers, and standards and guidelines. The inclusion of these sources was deemed essential to capture the regulatory landscape and industry best practices that shape the implementation of Industry 5.0 and human-machine collaboration frameworks.

After the corresponding iterations of study, additional sources were added, and some were excluded due to duplicate research or limited contribution to the document. The final list of references represents a balanced corpus spanning academic research, policy documents, and industry perspectives, providing a comprehensive view of Industry 5.0 and human-machine collaboration across theoretical, regulatory, and practical dimensions.

2. The Emergence of Industry 5.0: Beyond Automation

Industry 5.0 has emerged as a vision to go beyond automation-centric Industry 4.0 by reorienting industrial progress toward human-centric and sustainable goals [10]. While Industry 4.0 revolved around smart factories, cyber-physical systems, IoT, and AI to maximize efficiency and automation, Industry 5.0 shifts the paradigm back toward human-centricity, elevating workers from end-users to active collaborators. (see [1], This evolution recognizes that the next leap in productivity and innovation will come from collaboration between humans and machines, rather than automation in isolation [11]. In this paradigm, advanced technologies are seen as partners that enhance human capabilities, enabling personalized production and creative solutions that pure automation cannot achieve [12].

As Table 1 indicates, technology contributes by augmenting human abilities, yet it realizes its full potential only when paired with the explainability and interpretability obligations anchored in the human domain. Simultaneously, the ethics domain embeds accountability and transparency, ensuring that technology–human synergies do not drift into opaque or irresponsible practices.

Table 1. Core domains reinforcing human-centered AI in Industry 5.0.

Characteristic	Technology	Humans	Ethics
Focus	Augmenting human abilities	Algorithmic explainability and interpretability	Ethical, responsible systems
Goal	Enhance performance	Ensure effective human oversight	Embed ethical considerations
Key Aspect	Supportive, non-substitutive design	Maintain user confidence and control	Transparent and accountable governance

Taken together, the three columns illustrate that Industry 5.0's promise derives not from any single component but from the intersection of supportive design, effective oversight, and principled governance. This convergence is precisely what differentiates Industry 5.0 from its automation-centric predecessor and grounds our subsequent discussion of Trust by Design.

2.1. Pillars of Industry 5.0

The European Commission defines Industry 5.0 as a human-centered, sustainable, and resilient industrial paradigm [13,14]. These three pillars, shown in Figure 1, differentiate Industry 5.0 from its predecessor:

- Human-centricity: Fundamental human needs and well-being are prioritized in design and production processes. Rather than replacing humans, technology is used to empower workers, improve safety, and tailor production to individual needs. For example, even with increasing automation, human insight is valued for handling uncertainties and ensuring flexibility on the factory floor. This pillar is a compromise to uphold human dignity and foster meaningful work while countering fears of alienation by automation.
- Sustainability: Industry 5.0 emphasizes environmentally sustainable and circular production systems. This includes the achievement of carbon-neutral operations, prioritizing recycling and reusing resources, and minimizing waste. The objective is to align industrial growth with ecological responsibility, beyond Industry 4.0, by including societal and environmental value [15]. In practice, AI-optimized processes should not only save costs but also contemplate the reduction of energy consumption and emissions to contribute to achieving global sustainability goals.
- Resilience: To develop the capacity to withstand and adapt to crises—whether pandemics, economic disruptions, or supply chain shocks. Industry 5.0 prioritizes robust design and contingency planning so that critical infrastructure and supply lines continue to function under stress [16]. Technologies from Industry 4.0 (such as IoT sensors and predictive analytics) are leveraged with a new focus on risk mitigation and agility. The COVID-19 pandemic and other recent crises highlighted the need for such resilience and informed the Industry 5.0 agenda.

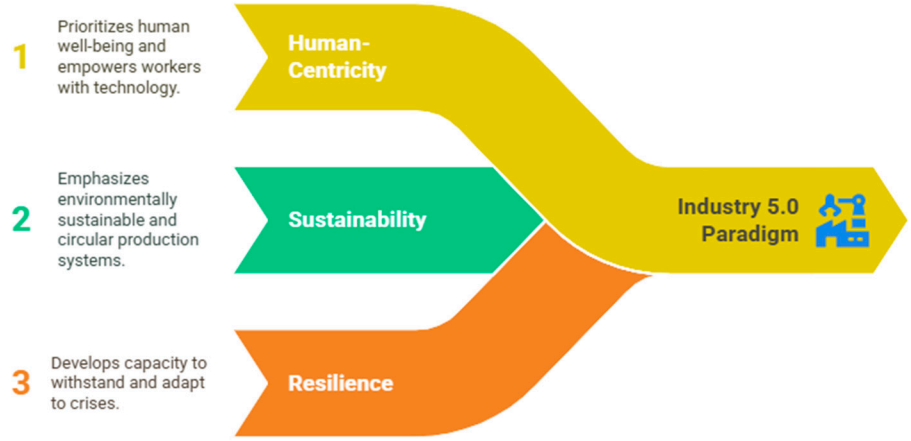

Figure 1. The three pillars of Industry 5.0—human-centricity (empowering workers and prioritizing well-being), sustainability (circular, eco-responsible production), and resilience (crisis-ready design and operations).

2.2. Relevance of Collaborative Intelligence

Crucially, collaborative intelligence ties these pillars together by enabling human–machine collaboration to meet both productivity and societal goals [6,17]. In this paradigm, humans

are not just end-users of automated systems but active collaborators with AI and robots. For instance, collaborative robots ("cobots") physically assist workers in manufacturing tasks, combining the robot's precision and strength with the human's problem-solving skills and adaptability [18]. Similarly, AI decision support systems augment human decision-makers, providing data-driven insights while the human applies contextual judgment. The aim is to harness the "best of both worlds"—human creativity and ethical judgment with machine speed and accuracy [13]. Industry 5.0 envisions factories where "people and smart machines work together in harmony" to co-create value, leading to individualized products, improved worker safety, and greater innovation [19].

Despite its promise, implementing Industry 5.0 faces significant challenges [20]. One challenge is technological integration: upgrading Industry 4.0 infrastructures to support seamless human-machine teaming and ensuring interoperability among diverse systems (robots, AI platforms, IoT devices). More sophisticated systems such as networked sensors, collaborative robots, and intelligent assistants are required to realize this vision, and companies may struggle with the complexity of integrating these into existing workflows.

Workforce readiness and skills are without a doubt another challenge; employees need training to effectively collaborate with AI systems, and there may be resistance to adopting unfamiliar technologies. Hassan et al. (2024) [21] noted that a lack of skilled staff and adequate training is a current barrier to Industry 5.0 adoption. However, organizational culture reflected by middle-management resistance can also impede adoption; studies found that employee and middle-management resistance significantly slowed Industry 4.0 implementations [22]. Industry 5.0 must overcome this trend by demonstrating clear value to workers and involving them in the transition [23]. Furthermore, businesses must invest in these new technologies at a time when the return on investment might be uncertain—high upfront costs and the drawn-out process of transformation are cited as ongoing issues [24].

Cybersecurity and privacy concerns add another layer of complexity. As Industry 5.0 connects more devices and relies on data sharing (including potentially sensitive personal data from wearables or AI monitoring of the worker), organizations must address GDPR risks or face breaches of trust. Indeed, a systematic risk analysis identified cybersecurity threats, data privacy issues, and ethical problems as key risk categories for Industry 5.0 systems. Without robust safeguards, increased connectivity could lead to vulnerabilities that undermine system resilience [25,26].

Due to the traditional drive for productivity inherent to industrialization, human-centric and sustainability goals are neglected and often overshadowed. This requires a mindset shift that introduces as metrics of success in Industry 5.0 the well-being of workers and environmental impact, not just throughput and profit [27]. This aligns with calls for stakeholder well-being as the "ultimate focus" of human–machine collaboration in Industry 5.0 [28].

In conclusion, Industry 5.0 provides a transformative approach that integrates human creativity and values into advanced industrial systems. It seeks to leverage collaborative intelligence to achieve a balanced blend of technological innovation with human-centric outcomes, sustainability, and resilience. Realizing this vision demands addressing implementation challenges through upskilling, organizational change management, robust security/privacy measures, and, importantly, developing ethical frameworks that ensure technology serves humanity's long-term interests. The next sections will discuss those ethical considerations and propose a framework to embed trust and ethics into Industry 5.0's collaborative intelligence systems from the ground up.

3. Ethical Considerations in Collaborative Intelligence

As humans and AI systems increasingly work side by side, trust becomes the linchpin of effective collaboration. In collaborative intelligence, trust is a person's willingness to rely on a machine under uncertainty, expecting it to act beneficially or at least acceptably [7,29]. It combines cognitive judgments (e.g., knowing the system's capabilities and track record) with an emotional readiness to be vulnerable to its actions [30]. Trust is dynamic; it grows or erodes through repeated interactions and is a double-edged sword: too little trust leads to underuse, while too much invites overreliance and costly failures.

Moreover, there is a considerable risk of critical-thinking erosion due to overreliance on AI. Users who come to depend too heavily on AI can fall prey to automation complacency—deferring even simple judgments to the machine. [31] shows that, in high-stakes domains such as medical training, overreliance on AI tutors not only masks system errors but also undermines practitioners' own critical-thinking skills, leading to a "deskilling" effect when the AI is unavailable or mistaken. To counter this, Trust by Design embeds calibrated override points (mandatory human confirmation for critical actions) and "failure-mode drills", where users deliberately confront AI errors so they learn to detect and correct them. Thus, designing collaborative intelligence systems demands calibrated trust: justified confidence rather than blind faith.

3.1. Ethical Challenges to Advance Collaborative Intelligence

As illustrated in Figure 2, several ethical challenges must be addressed to achieve trust and ensure these human-machine collaborations uphold broader societal values:

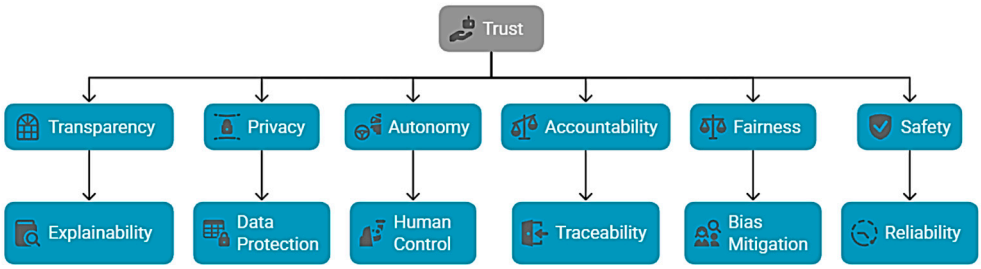

Figure 2. Ethical considerations in collaborative intelligence: foundational principles of trust.

- Transparency: Collaborative AI systems often operate as "black boxes" that are difficult for humans to understand. Lack of transparency in how an AI makes decisions or how a robot's actions are determined can breed mistrust and confusion. Ethically, there is a demand for explainability—humans should be able to get clear, intelligible reasons for an AI system's recommendations or actions [32,33]. Transparency is also critical for informed consent: workers should know what data is being collected and how algorithms are using it. Without transparency, power imbalances emerge where only the system (or its vendors) "know" why certain decisions are made, leaving users in the dark. Ensuring transparency (through user dashboards, visualizations, or explainable AI techniques) can build understanding and calibrated trust, as users can verify and make sense of the system's behavior [34].
- Privacy: Collaborative intelligence systems frequently rely on large amounts of data—including personal and sensitive data about workers that can be obtained from wearable devices and smart tools, which are an embodiment of the Internet of Things (IoT), which can later be used to train different AI models to estimate workers' performance, health indicators, or movements, converging towards what can be referred to as Artificial Intelligence of Things (AIoT) [35]. This raises concerns about data privacy and surveillance. If not properly governed, such data collection can infringe

on workers' privacy rights and create a climate of surveillance that erodes trust and autonomy [36]. Ethical use of collaborative systems demands strict adherence to privacy principles: data should be collected only for legitimate, agreed-upon purposes and with consent wherever possible; it should be anonymized or minimized to protect identities; and robust cybersecurity must protect it from breaches. Privacy considerations extend beyond the workplace—as collaborative robots interact in shared human environments, video or sensor data could inadvertently capture bystanders or sensitive information, necessitating careful privacy-by-design measures.

- Autonomy and human agency: A core ethical tension in human-machine collaboration is balancing machine autonomy with human control. On one hand, the AI or robot needs a degree of autonomy to be useful (e.g., a cobot adjusting its movements in real-time or an AI filtering relevant information). On the other hand, if the machine's autonomy encroaches on human decision-making without oversight, it can diminish human agency and accountability. Who is in charge? Ethically, humans should retain meaningful control over the overall task and have the ability to overrule or adjust the machine's actions according to the six possible paradigms of Human-Machine Interaction: Humans in the Loop (HITL), Humans on the Loop (HOTL), Humans out of the Loop (HOOTL), Humans alongside the Loop (HATL), Humans-in-command (HIC), and Coactive Systems [37]. Maintaining human agency is not just about operational control but also psychological empowerment [38]—workers should feel they are active participants, not passive servants to an AI's instructions.
- Accountability: With shared human–AI decision-making, it can become unclear who is accountable when something goes wrong. Is it the worker using the AI, the AI's developer, the employer deploying it, or the machine itself (which, lacking personhood, cannot bear responsibility in a moral or legal sense)? This diffusion of responsibility is a serious ethical and legal challenge. Collaborative systems should be designed such that accountability is traceable and assignable [39]—for instance, by keeping logs of AI decisions, providing tools for audit, and defining roles so that humans have specific oversight duties. If an AI system recommends a faulty course of action, there should be mechanisms to investigate whether the human followed blindly or whether the AI provided misleading information. Ethically, companies and technology providers need to share accountability by ensuring proper training, setting reasonable expectations for human intervention, and responding to incidents with transparency. In the absence of clear accountability, trust in the system will erode—people will be reluctant to use systems if they fear being scapegoated for their errors, or conversely if they worry no one will be responsible if the system harms them.
- Fairness and non-discrimination are of particular ethical consideration in collaborative intelligence. AI systems embedded in industrial settings might make decisions about task assignments, evaluations of work quality, or even hiring and promotion (in advanced scenarios). If these algorithms carry biases, they could unfairly disadvantage certain groups [40]. For example, an AI scheduling system might inadvertently assign more repetitive or risky tasks to certain workers based on biased data, or a decision support tool might underrate the contributions of older workers if it is not designed carefully [41]. Ensuring fairness requires careful design and continual monitoring of algorithms to detect disparate impacts. It also intersects with diversity and inclusion—a human-centric Industry 5.0 must accommodate diverse needs and avoid one-size-fits-all automation that ignores, for instance, workers with disabilities or different skill profiles. Engaging a diverse range of stakeholders in system design can help pre-empt bias and foster equity.

- Safety and reliability: In collaborative environments, physical and psychological safety is paramount. Ethically, robots and AI should be rigorously tested to fail-safe—meaning any malfunction should default to a safe state that minimizes risk to humans. The ISO and industry safety standards (such as ISO 10218-1:2025, ISO 10218-2:2025, and ISO/TS 15066:2016 for robot safety and collaborative robots [42–44]) provide guidelines for physical robot collaboration limits (such as force and speed limits when near humans). However, beyond physical safety, there are psychosocial safety concerns: research has found that the introduction of cobots can cause stress, job insecurity, and role ambiguity for workers if not handled properly [45]. These manifest as psychosocial hazards that can affect mental health. Ethical deployment requires addressing such safety holistically by providing training to build confidence, ensuring the technology truly reduces (and does not add to) cognitive workload, and maintaining a work environment where humans feel safe working with and alongside robots. Reliability of AI is equally crucial; frequent errors or unpredictable behavior quickly destroy trust. Thus, an ethical system must not promise more than it can deliver—transparency about the system's limits and uncertainties is better than a misleading aura of infallibility.

3.2. Stakeholder Perspectives on Ethical Challenges

These challenges, as depicted in Figure 3, must be viewed from multiple stakeholder perspectives to be fully understood:

Figure 3. Stakeholder dynamics in Industry 5.0: a multi-dimensional view.

- Workers (human operators): Front-line workers are directly affected by collaborative systems, with job security a primary concern. Collaborative robots and AI can provoke fears of displacement or role downgrading; studies show workers often view cobots as threats, particularly when collaboration seems minimal and replacement plausible [46]. Resistance stems from perceived threats to autonomy, skill obsolescence, and safety, so organizations must offer transparency, training, and dialogue to ease AI anxiety [47]. Such fears can breed stress, erode trust, and raise safety (e.g., "Will the robot strike me?"), agency ("Do I still control my work?"), and privacy ("Is constant monitoring invasive?") worries. Ethically, workers expect respectful treatment and prioritization of their well-being. If a collaborative system demonstrably reduces drudgery or injury risk, and management clearly communicates its benefits, workers are likelier to accept it. Involving workers in design and rollout through participatory design, training sessions, and feedback loops is widely recommended to address their concerns [48].
- Employers (organizations/managers): Employers seek productivity gains, quality improvements, and flexibility from collaborative intelligence. They are stakeholders in ensuring ROI on these technologies. However, they also carry responsibilities for

worker safety, legal compliance, and maintaining a motivated workforce. Ethically, employers must balance profit motives with the duty of care for employees. They may worry about liability—if an AI causes a bad decision, the company could be responsible. Thus, they have an interest in clear accountability frameworks and reliable system performance. Change management is another concern: how to implement collaborative systems without disrupting operations or sparking labor disputes. From a trust perspective, employers need to build organizational trust—workers must trust that management is introducing AI/robots to assist rather than surveil or replace them. Research suggests that engaging employees early and transparently can smooth the transition and reduce psychosocial risks. Employers also must consider skill development—they should provide training so employees can effectively collaborate with AI, which in turn can improve acceptance and outcomes [21]. Forward-looking employers see collaborative intelligence as augmenting their human talent, not depreciating it.

- Technology providers (engineers and vendors): Those who design and supply collaborative AI/robot systems have a stakeholder interest in the successful and ethical use of their products. Their reputation and market success may depend on users trusting their technology. Providers face the challenge of translating ethical principles into design features—for example, building explainability, user-friendly interfaces, and safety mechanisms. Many tech companies are now adopting "responsible AI" charters, recognizing that neglecting ethics can lead to user backlash or regulatory action [49]. Providers might worry about intellectual property vs. transparency—how much of their algorithm's inner workings to reveal. Ethically, they have a responsibility to ensure their systems are not biased or dangerous, which requires thorough testing and perhaps adhering to standards or certifications. There is also the issue of support and updates: a collaborative system may evolve with software updates or new data; providers should continuously monitor for ethical or safety issues post-deployment (sometimes in collaboration with the client). In essence, technology providers must practice Ethics by Design and often need to educate and support their clients in deploying technology in line with ethical best practices.
- Society and regulators: Society has a direct stake in Industry 5.0's trajectory—it will shape employment, inequality, and well-being. Many hope it delivers meaningful jobs and sustainable practices, not just greater output [1]. he public cares whether collaborative intelligence augments workers (upskilling and safer roles) or merely eliminates positions and increases surveillance. This raises justice issues: ensuring productivity gains translate into better conditions or work-life balance, not solely corporate profits. Regulators and policymakers (see Section 7) are crafting guidelines stressing human oversight, non-discrimination, and privacy—for example, the EU's Trustworthy AI framework mandates human agency, robustness, privacy, transparency, diversity, and accountability [50,51]. If collaborative systems violate fundamental rights or societal values, there could be regulatory penalties or public pushback (for instance, strong unions might oppose dangerous or dehumanizing tech). Society also includes consumers: in some settings (such as healthcare or customer service), the end-users of collaborative intelligence outputs are the public, who will trust a company more if they know its AI is ethically governed. Overall, societal stakeholders demand that Industry 5.0's trajectory align with the public interest—creating inclusive, safe, and human-centered progress rather than exacerbating social harms.

3.3. Alignment and Limits of Current Frameworks Towards Industry 5.0

Current ethical approaches to AI and automation (such as high-level principles from bodies such as the EU or IEEE) provide important foundations but also have limitations. Many organizations have adopted ethics charters enumerating values such as transparency, fairness, and accountability. However, the numbers are still low, as many organizations are still catching up with these new guidelines [52]. Principles alone do not guarantee practice—there is frequently a gap between stating "we value privacy" and operationalizing privacy in system architectures.

From our literature review, fourteen recent studies on ethics, trust, and human-centricity in Industry 5.0 were identified. Table 2 summarizes each work's approach, domain focus, core objective, methodology type, ethics-and-trust emphasis, and alignment to the three Industry 5.0 pillars (human-centricity, sustainability, and resilience).

Table 2. Recent studies on ethics, trust, and human-centricity in Industry 5.0 (2022–2025).

Paper	Approach	Domain Focus	Core Objective	Proposed Methodology	Ethics and Trust Focus	Industry 5.0 Alignment
Bohr (2025) [52]	Case study narrative of adopting IEEE 7000 ethics-by-design standard.	Ethics-by-design in software engineering.	Share lessons learned in translating ethical values into system requirements.	Risk-based integration of IEEE 7000	Bridges high-level values to concrete requirements; emphasizes traceability and stakeholder engagement.	Governance framework for ethics-by-design in I5.0 development
Brey and Dainow (2024) [53]	Conceptual development of the EbD-AI ethics-by-design approach.	Ethically guided AI system design.	Present and compare a full EbD-AI framework adopted by EU Horizon ethics review.	Six-stage procedure: Values → stakeholder Review → monitoring	Comprehensive integration of seven ethics requirements; trust via upfront ethics embedding and review.	Practical methodology for embedding ethics in I5.0 AI lifecycles
Callari et al. (2024) [54]	Delphi-based co-creation of an ethical H-R collaboration framework.	Ethics in human–robot collaboration for people-centric manufacturing.	Co-design, with experts, a holistic ethical framework at shop floor, organizational, and societal levels.	Three-round Delphi with ethics experts	Central governance for ethics awareness, responsibility, and accountability to foster trust.	Human-centric pillar; governance for responsible robotics integration
Fraga-Lamas et al. [14]	Analytical review of blockchain's role in Industry 5.0.	Blockchain for human-centric, sustainable, and resilient applications.	Provide a detailed guide on how blockchain can underpin I5.0's pillars and what design factors to consider.	Taxonomy by I5.0 pillar + design guidelines	Positions blockchain as a trust anchor (immutability, decentralization); ethical focus on worker empowerment and data sovereignty.	Supports all three I5.0 pillars via trustworthy data sharing
Ghobakhloo et al. (2024) [27]	Content synthesis + HF-ISM roadmap modeling.	Roadmap from I4.0 digital manufacturing to I5.0 digital society.	Clarify drivers behind Industry 5.0's emergence and sequence I4.0 sustainability functions to enable I5.0 goals.	I4.0 sustainability synthesis + HF-ISM for function interde-pendencies	Emphasizes governance, socio-environmental sustainability and trust via stakeholder-driven digitalization.	Integrates economic, social, and environmental pillars; resilience roadmap
Langås et al. (2025) [19]	Integrative review and conceptual mapping of HRT, digital twins, and ML synergy.	Sustainable manufacturing through human–robot teaming and digital twins.	Examine how combining HRT, DT, and ML can enable safe, efficient, and sustainable human-centric production.	HRI → HRC → pHRC → HRT mapping with DT/ML enablers	Implicit ethical concern for worker safety and well-being; limited explicit treatment of fairness.	Aligns digital-physical integration with human-centric and sustainability pillars
Martini et al. (2024) [13]	Position paper on HCAI in I5.0 and circular economy.	Human-centered AI and circular economy in additive manufacturing.	Identify major challenges and prospective areas for human-centered AI in I5.0.	Mapping HCAI onto AM workflows + policy analysis	Emphasizes ethics, transparency and regulation for HCAI; trust via participatory design.	Human-centric and sustainability pillars; circular economy enabler
Palumbo et al. (2024) [55]	SLR of objective metrics for ethical AI aligned with EU Trustworthy AI.	AI ethics metrics per seven EU principles.	Identify and categorize objective metrics to assess AI Ethics.	SLR protocol mapping metrics to 7 ethics principles	Deep focus on metrics for fairness, transparency, accountability; trust via measurable compliance.	Provides measurable KPIs for embedding ethics in I5.0 systems

Table 2. *Cont.*

Paper	Approach	Domain Focus	Core Objective	Proposed Methodology	Ethics and Trust Focus	Industry 5.0 Alignment
Przegalińska et al. (2025) [6]	Experimental evaluation of generative AI as a collaborative assistant.	Human–AI collaboration in workplace tasks.	Explore how generative AI tools optimize organizational task performance across complexity and creativity.	RBV + TTF task typology + live generative-AI experiments	Supports trust by showing AI's positive sentiment and clarity; ethics of augmentation, not replacement.	Human-AI teaming; hybrid intelligence; organizational performance
Riar et al. (2025) [9]	Experimental comparison of three design interventions (non-gamified, gameful, playful) in VR.	Human–robot collaboration (HRC) trust-building via gamification.	Investigate how gameful versus playful design influences cognitive and affective trust in collaborative robots.	The three-arm VR experiment manipulates gamification archetypes and measures trust outcomes.	Directly targets affective trust and specific antecedents; ethics in ensuring positive emotional connection.	Emphasizes user experience in cobots; human-centric interaction
Santos et al. (2024) [25]	Analytical review of cyberattack surfaces and countermeasures; critical analysis of existing frameworks.	Cybersecurity within Industry 5.0.	Identify new threats posed by I5.0 enabling technologies and evaluate current industrial implementation frameworks to secure the transition from I4.0 to I5.0.	Threat matrix + I4.0 framework gap analysis	Emphasizes safeguarding human-centric values, privacy and mental health by robust cybersecurity; builds trust through resilience.	Highlights resilience and human-centricity via robust cybersecurity
Textor et al. (2022) [32]	Mixed-methods exploration of ethics in human–AI teams.	Ethics and trust dynamics in human–AI teaming.	Uncover how ethical considerations shape—and are shaped by—trust in collaborative AI settings.	Interviews + surveys	Core focus on the co-dependence of ethics and trust; transparency and accountability emerge as key.	Underpins ethical governance in human-AI collaboration
Thurzo (2025) [56]	Architectural design of a provable-ethics "ethical firewall".	Provable ethics and explainability in high-stakes AI (medical/educational).	Embed mathematically verifiable ethical constraints into AI decision cores.	Formal logic + blockchain + Bayesian escalation	Ethics and trust engineered into AI core—decisions provably aligned with human values.	Ensures real-time transparency and accountability in high-stakes I5.0 AI
Trstenjak et al. (2025) [2]	SLR of human factors and ergonomics in Industry 5.0 work environments.	Identify characteristics, dimensions, and principles enabling/hindering human-centric work designs.	PRISMA SLR of WoS (983 records → 119); thematic grouping into nine ergonomics domains.	PRISMA WoS review into nine domains	Addresses I5.0's human-centric pillar by detailing ergonomic requirements for collaborative work.	Human-centric socio-technical design; sustainability; resilience

Among the reviewed studies, only a handful explicitly engage all three core pillars of Industry 5.0: human-centricity, sustainability, and resilience. In particular, the roadmap study in [27] stands out for its system-level treatment, mapping nine Industry 4.0 functions through a content-centric synthesis to show how automation, circularity, and real-time integration can be sequenced not only to empower humans (human-centricity) but also to deliver environmental and social sustainability while aiming to build antifragile, resilient business models.

By contrast, most experimental and framework papers address only one or two pillars in depth. Riar et al. and Textor et al. [9,32], for instance, deepen our empirical understanding of human-centric trust dynamics but do not directly tackle environmental or organizational resilience. Martini and Bellisario's position piece on human-centered AI in additive manufacturing brings human-centric and sustainability concerns into dialogue but leaves resilience as an implicit, rather than explicit, design objective. Similarly, Brey and Dainow's Ethics by Design [53] and Bohr's account of the IEEE 7000 implementation journey [52] operationalize human agency and system robustness (resilience), yet do not engage directly with sustainability metrics or eco-social equilibria.

Emergent themes around trust and ethics reveal both progress and persistent gaps. Quantitative studies identify conditions under which ethical violations diminish trust—yet also uncover nuanced cases in which trust may decouple from perceived ethicality, suggesting that AI purpose and process models warrant closer scrutiny. Palumbo et al. in [55]

note the scarcity of objective, quantifiable metrics for most trustworthy AI principles beyond fairness, leaving developers without clear performance guardrails for human agency, transparency, or environmental well-being. Thurzo's "Ethical Firewall" architecture [56] offers an ambitious technical blueprint for embedding mathematically provable ethical constraints into AI decision systems, but it also surfaces significant performance and complexity overheads when layering formal verification and cryptographic immutability atop dynamic, learning-based agents.

The blockchain overview presented by Fraga-Lamas et al. in [14] interrogates technological trust and resilience in decentralized ledgers and highlights sustainability trade-offs in IIoT deployments, yet pays comparatively less attention to the lived, human-centered experience on the shop floor. Przegalińska et al.'s [6] and Callari et al.'s [54] human–AI collaboration and ethical framework studies address human-centric and organizational resilience dimensions (through improved decision-making and accountability) but only gesturally point toward broader socio-environmental sustainability outcomes. Finally, Trstenjak et al.'s ergonomics review in [2] maps safety, well-being, and cognitive workload considerations but again does not systematically link these to environmental or societal sustainability imperatives.

In sum, while every paper contributes valuable insights to one or two pillars of Industry 5.0, our proposed Trust by Design framework seeks to fill this gap by providing a structured, end-to-end approach that embeds trustworthiness and ethical safeguards into every stage of an AI system's lifecycle, aiming to ensure these technologies genuinely empower human collaborators, preserve their autonomy, and uphold the fairness, transparency, and resilience that all stakeholders require under the three pillars of Industry 5.0.

This synthesis directly answers our first research question by identifying the unique ethical challenges of human-centric collaborative intelligence in Industry 5.0 and highlighting where existing work falls short.

4. Proposed Ethical Framework: Trust by Design

4.1. Foundational Principles of Trust by Design

Trust by Design is a framework of principles and practices intended to ensure that collaborative intelligence systems are developed and deployed in ways that inherently foster trustworthiness and ethical behavior, in line with our second research question. The framework, whose core principles are outlined in Figure 4, is inspired by the concept of "Ethics by Design", integrating ethical reasoning capabilities and considerations into technology from the earliest stages [53,57], but focuses specifically on building trust as the outcome of ethical alignment. The core idea is that trust is not an afterthought or something to be addressed only by training users; rather, trust must be "designed into" the system's architecture, user experience, and governance. Below we outline the core principles and values that constitute the foundation of Trust by Design, followed by the practical models and mechanisms that operationalize these principles.

Core Principles of Trust by Design:

1. Human agency and empowerment: The system must augment rather than replace human intelligence, preserving human control where it matters. Collaborative AI should be designed to enhance human capabilities (cognitive or physical) and support human decision-making while ensuring that users can override or steer the AI's actions when necessary. This principle upholds the value of autonomy—the human operator remains an active agent in the loop. For example, an AI decision aid might present options and recommendations but let the human confirm or adjust the final decision, thus acting as a supportive colleague, not an infallible oracle. All design

choices (from default settings to emergency stop buttons) should reinforce that the human is ultimately in command.

2. Transparency and explainability: The system should function as a "glass box" to the extent feasible, providing clear explanations or insights into its operations [32]. This includes making the AI's decision logic interpretable and the robot's intent foreseeable (through visual signals or predictable motions in the case of physical cobots). When users understand why the AI produced a certain output or what the robot is about to do, they can develop informed trust [7]. Trust by Design calls for integrating explainability features (e.g., justification dialogues, user queries to the AI) and ensuring the UI communicates uncertainty or confidence levels of the AI. Even if the underlying algorithms are complex (such as deep learning), the system should translate that complexity into user-relevant terms (such as highlighting which factors most influenced a recommendation). Transparency extends to data practices—users should know what data is being collected and how it is used (akin to a privacy notice embedded in the interface).
3. Privacy and data governance: From the outset, systems should adhere to "privacy by design" [58–60]—collecting minimal data, securing it rigorously, and using it in ethically and legally appropriate ways. In Trust by Design, any personal or sensitive data (e.g., worker biometrics, productivity metrics) is handled with confidentiality and respect for user consent. Technical measures such as encryption, access controls, and on-device processing (to avoid unnecessary data transmission) are employed to protect privacy. Additionally, the framework mandates transparency to users about data usage and provides options to opt out or control certain data sharing where possible. By safeguarding privacy, the system demonstrates respect for the user, which is fundamental for trust.
4. Fairness and inclusivity: The framework embeds checks to ensure the system's decisions or actions do not systematically disadvantage any individual or group without justification. This involves using bias mitigation techniques during model training (for AI components) and diverse user testing to see how the system performs across different scenarios and users. The values of equality and justice require that, for instance, a decision support AI should apply the same standards to everyone and be audited for bias. Research has shown that AI-enabled recruitment tools can perpetuate biases, leading to discriminatory hiring practices based on gender, race, or other characteristics [61]. Such patterns should be detected and corrected to ensure fair treatment of all candidates. Inclusivity also means designing the human interface with accessibility in mind (for different physical abilities, language skills, etc.), ensuring all workers can effectively collaborate with the system.
5. Safety and reliability: Borrowing from the principle of technical robustness in trustworthy AI [51], Trust by Design prioritizes safety measures at all levels. Physical safety is addressed by compliant robot design, safe stop mechanisms, and strict testing against scenarios of potential collision or misuse. The system should be fail-safe and fail-transparent—if errors occur, the system defaults to a safe state and informs the user. Reliability entails thorough validation so that the system behaves predictably within its defined operating conditions. This principle builds trust by minimizing the occurrence of unexpected or dangerous behavior. Furthermore, psychosocial safety is included: features or policies are in place to mitigate stress (for example, the system might be designed to adapt to the user's pace rather than enforcing an uncomfortable speed). Alarms or notifications are tuned to avoid causing alarm fatigue or distraction. Overall, the system's robust performance and safety track record form the bedrock of user confidence.

6. Accountability and auditability: The design should allow for tracing decisions and actions back to their source. This means maintaining logs of AI recommendations, robot actions, and human overrides in a secure but reviewable manner. In case of an incident or ethical dilemma, these records enable an audit to understand what happened and why. More proactively, the system can include self-checks or ethical governors—for instance, an AI could have constraints that prevent it from recommending actions violating certain rules (much like how a thermostat would not go beyond certain limits). Accountability is also organizational: roles are defined so that there is always a human responsible for monitoring the system's outputs (e.g., a shift supervisor who reviews all critical AI suggestions). This clarity prevents diffusion of responsibility and assures users that the system is under accountable oversight.
7. User involvement and training: While more of a process principle than a design element, Trust by Design emphasizes co-design with end-users and comprehensive training as part of system development. By involving workers and domain experts in the design phase (through feedback sessions, pilots, etc.), designers can capture contextual ethical issues and trust concerns early on. The framework treats user education as part of the design: intuitive tutorials, simulations, and continuous learning resources are built into the rollout so that users gain competence and confidence in interacting with the AI/robot. A system is trustworthy not only because of its internal qualities but because users feel competent in using it; thus, designing the learning curve and support materials is an ethical imperative here.

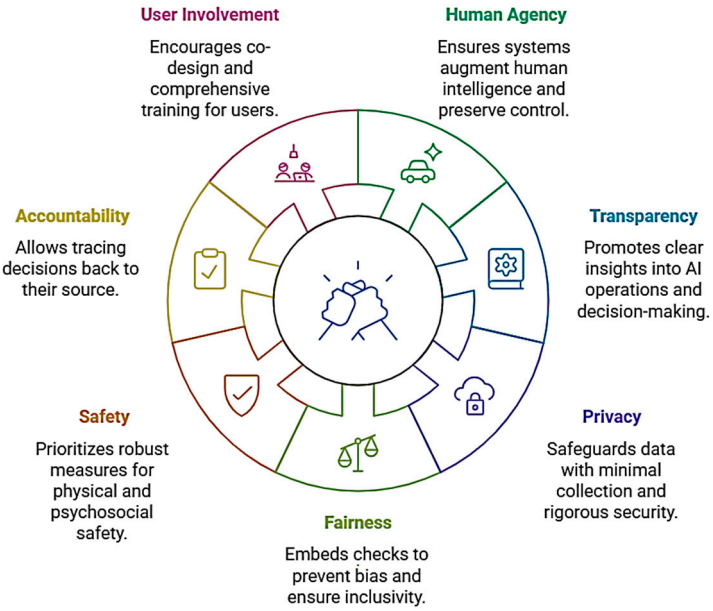

Figure 4. Trust by design framework: core principles.

These core principles align closely with high-level frameworks like the EU's trustworthy AI guidelines (human agency, transparency, etc.), but Trust by Design tailors them to collaborative intelligence and provides actionable interpretation for industrial contexts. Next, we describe ethical decision-making models and trust-building mechanisms that implement these principles throughout the system lifecycle.

Ethical decision-making models for collaborative systems: To embed ethics into the behavior of AI components, Trust by Design can leverage models such as Value Sensitive Design (VSD) [62] and multi-objective decision frameworks that include ethical utilities [55]. Value Sensitive Design, for example, provides a methodology to systematically consider human values (such as safety, autonomy, and privacy) during the design process by in-

volving stakeholders and iteratively refining the system to address value tensions [54,63]. In a collaborative robot scenario, VSD might involve interviewing workers about what would make them trust the robot and then designing features in response (such as a pause function or certain courteous behaviors from the robot).

Another approach is incorporating ethical reasoning modules [64] with a focus on AI. For instance, a collaborative AI could be equipped with a simple rule-based system that checks its recommendations against ethical constraints—akin to a mini conscience. If an AI planner for a factory schedule finds an optimization that boosts output but causes excessive workload on one person, an ethical rule could flag this as violating fairness and prompt the AI to seek an alternative solution. Researchers in AI ethics have explored techniques such as constrained optimization and utility functions augmented with fairness or safety terms so the AI intrinsically balances performance with ethical considerations. Trust by Design advocates for such ethics-aware algorithms.

From a theoretical standpoint, designers should consider classical ethical theories as lenses: utilitarian thinking to ensure overall well-being is increased (but tempered so that it does not justify harming a minority for greater good), deontological rules to respect fundamental rights (such as "do not deceive the user" and "do not violate privacy"), and virtue ethics by encouraging practices that cultivate trustworthiness (such as honesty and reliability). In practice, this might mean hard-coding certain constraints (deontological) and using system metrics that reflect collective benefit (utilitarian) while fostering a company culture where engineers aim to be virtuous practitioners (virtue ethics). By blending these, the framework ensures no single ethical lens dominates to the detriment of others, creating a more robust ethical decision-making approach.

4.2. Embedding Trust in the Lifecycle of AI Systems

As depicted in Figure 5, Trust by Design is not a one-time checklist but a lifecycle approach. Trust is built at each stage, from initial design and development through testing, deployment, operation, and ongoing feedback. This cyclical framework ensures that ethical considerations and trust-building are continuous and iterative processes, rather than static checkpoints.

1. Design and development stage: During design, user research and risk analysis inform features that directly address trust issues (e.g., adding an explanation panel after finding in user studies that operators mistrust opaque AI outputs). Simulation and modeling are used to foresee interaction patterns—for instance, simulate scenarios where the AI is wrong and ensure the system handles it gracefully (alerting the user, offering fallback). At this stage, ethical risk assessment is carried out to identify where things could go ethically wrong and to mitigate those risks upfront. Engineers incorporate redundant safety mechanisms so that if one component fails, another catches it (increasing reliability trust). Agile development methods can integrate ethics by having "ethical user stories"—e.g., "As a worker, I want to know why the scheduling AI gave me more shifts than my colleague, so that I feel the process is fair". This user story would lead to implementing an explanation or adjustment feature.
2. Testing and validation stage: The system is evaluated not only for functionality but also for ethical compliance and trust factors. This might involve user testing sessions specifically to gauge trust: do users feel comfortable after using the system? Can they correctly recount what the AI or robot did and why? Any confusion or discomfort is a red flag to address. Measures such as the Trust Scale [65] (a survey instrument from human factors research) can quantify user trust levels during trials. Additionally, safety tests under various edge cases show whether the system meets the safety principle. If during testing a scenario reveals, say, an ambiguous robot motion that

startles workers, designers refine the motion planning to be more transparent (maybe slowing down and using a signal when humans approach). The system may also undergo an ethical audit by an internal or external committee to verify that privacy controls work, data bias is absent, and so on. By iterating at this stage, the final product that goes live is already tuned for trustworthiness.

3. Deployment stage: Initial deployment is performed in a pilot or phased manner to build trust gradually. Trust by Design encourages introducing the system with a human-led orientation: explaining to the team the goals of the system, how it works (in lay terms), and addressing questions openly. Early positive experiences are crucial—thus, maybe the system starts with assisting in low-stakes tasks and, as users gain confidence, moves to more critical functions. Mentorship models can be employed (a tech champion on the factory floor helps peers learn the system, building peer trust). Moreover, the system itself can have built-in tutorials or AI-guided onboarding—for example, a collaborative robot might initially operate in a slower "training mode" around new users, essentially earning trust by demonstrating consistent safe behavior, and only later ramp up to full speed.
4. Operation and maintenance stage: Trust is maintained through ongoing support and system transparency. The framework suggests continuous monitoring of system performance and user feedback. Dashboards for supervisors could show system health and any anomalies (transparency at the management level ensures they trust it and will advocate its use to workers). If the AI encounters a situation outside its training (novel input), it can either abstain or seek human confirmation, rather than act unpredictably—this humility in AI behavior (knowing when to defer to humans) significantly boosts trust. Regular training refreshers or update notes keep users in the loop on any changes, so they never feel the system is drifting beyond their understanding. From a technical side, predictive maintenance for hardware and retraining of AI models ensure the system remains reliable and up-to-date, preventing degradation of trust due to aging components or stale data.
5. Feedback and evolution stage: A trustworthy system welcomes user feedback and adapts. Trust by Design embeds feedback channels—perhaps a feature for users to flag if an AI suggestion did not seem right or if they felt uncomfortable at any point. This feedback is reviewed by the development team or ethics committee to identify new ethical issues or needed improvements. In effect, the ethical framework itself evolves: maybe real-world use uncovers a scenario not anticipated (e.g., workers developing an overreliance on the AI for trivial decisions). The organization can then tweak procedures or the system (such as adding periodic "are you sure?" prompts or rotating tasks to keep skills sharp) to correct this. This responsiveness shows users that their trust in the system and the organization is reciprocated—the company is committed to continuous ethical improvement, not just a one-off deployment.

By threading these trust-building mechanisms through the lifecycle, the Trust by Design approach creates a virtuous cycle: a system that is transparently and accountably designed engenders initial trust; proper training and safe initial experiences reinforce that trust; and continued reliability and responsiveness sustain it. Importantly, this approach recognizes trust as an ongoing relationship, not a static attribute. As noted by Merritt and Ilgen (2008) [66], trust in automation evolves with ongoing interactions, being influenced by the user's direct experiences and the system's demonstrated behavior. Therefore, our framework is not "once trustworthy, always trustworthy"—it includes provisions for verification and validation of ethical compliance throughout the system's life.

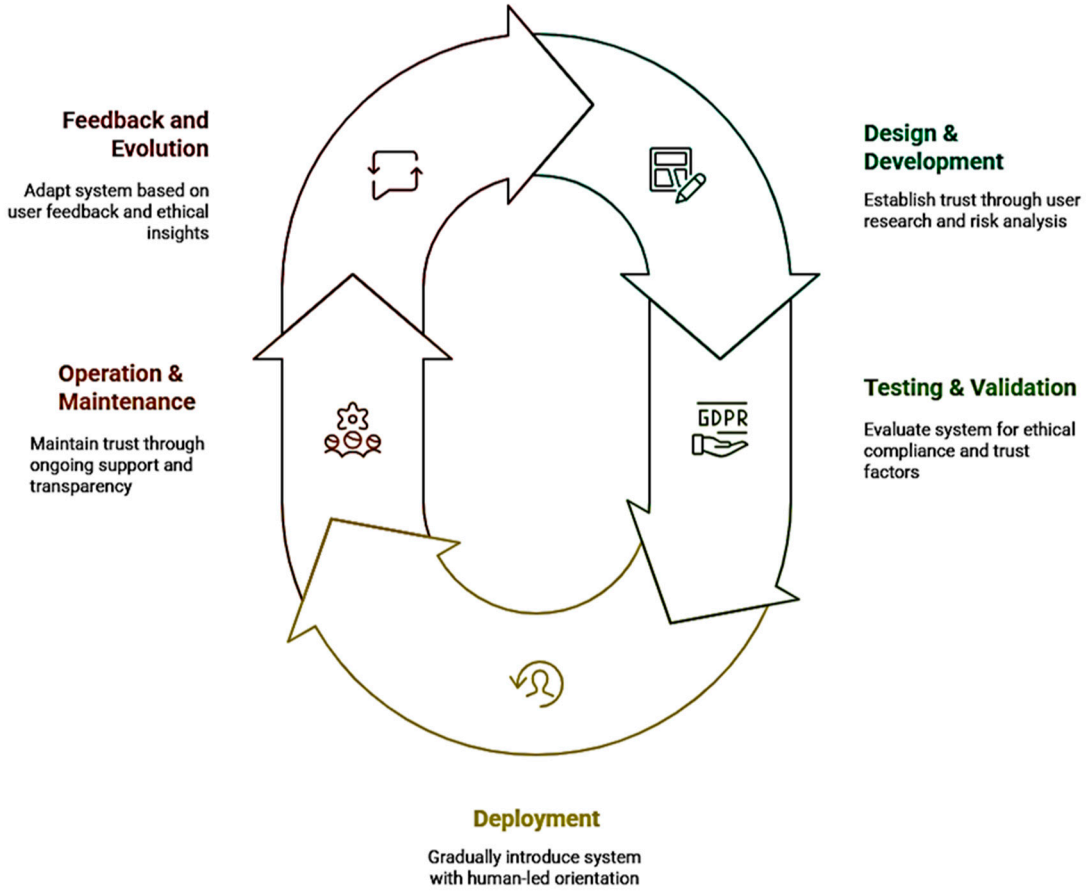

Figure 5. Trust-building lifecycle in AI systems.

Verification and validation of ethical compliance: To ensure the system truly adheres to the above principles, formal verification steps are included. This can involve ethical checklists such as the EU's ALTAI (Assessment List for Trustworthy AI) adapted for collaborative systems—a series of questions that developers and deployers answer and document, such as "Have we informed users of the system's purpose and limitations?" or "Have we tested for any bias in task allocation by the AI?" An internal or third-party audit could review these items and perhaps simulate adversarial scenarios (to test privacy or security). Some organizations establish an AI Ethics Review Board to sign off on new deployments, similar to an Institutional Review Board in research ethics, ensuring an independent perspective on compliance.

For physical systems, validation might include certification to standards (if available, e.g., an IEEE 7000-2021 certification for ethics-informed design process). Additionally, user acceptance tests that include ethical criteria (users must agree that "I felt in control" and "the system was transparent to me") act as validation from the stakeholder perspective.

Finally, Trust by Design encourages publishing certain aspects of the system's design or results of ethical audits (within IP limits) to stakeholders or even the public. This transparency about the ethical validation process itself can strengthen trust—users and society see that the system's creators have nothing to hide and have rigorously checked its trustworthiness.

4.3. Navigating Trade-Offs and Proportionality

While these core principles provide a foundation for trustworthy systems, real-world deployments may occasionally confront. Potential conflicts include:

- Transparency vs. privacy—explaining decisions without over-exposing personal data.

- Human agency vs. safety and reliability—letting people override the system without undermining safeguards.
- Accountability vs. privacy—keeping audit trails while respecting individual confidentiality.
- Fairness and inclusivity vs. safety and reliability—tailoring for diverse users without weakening predictable behavior.
- Autonomy vs. Fairness—giving supervisors discretion without reintroducing bias.

This can be tackled through a three-step proportionality heuristic:

(a) Identify stakeholders and values in tension: Clearly identify all stakeholders affected by the conflict and explicitly state the ethical principles or values that appear to be competing.
(b) Conduct relative impact analysis: Evaluate the potential impact and consequences of favoring each conflicting principle by applying tools such as a least-intrusion test, risk-benefit matrices, or ethical impact assessments.
(c) Mitigate and document chosen safeguards: Develop and select specific technical or procedural safeguards designed to harmonize the conflicting ethical requirements, seeking the least intrusive yet effective solutions. Document thoroughly the rationale, justifications, and chosen mitigations to ensure accountability, transparency, and auditability.

In a hypothetical scenario, our decision support AI must explain why it flags a worker's fatigue risk. Revealing the raw heart-rate trace would breach privacy, yet showing no rationale undermines transparency. The heuristic yields a middle path: we surface aggregated fatigue indicators and feature-importance rankings—demonstrating causality—while encrypting raw biometrics and limiting access to authorized medical staff. The decision log must record the trade-off analysis for future audits.

In summary, Trust by Design provides a holistic framework where ethical principles guide design, development incorporates trust-centric features, and verification mechanisms ensure those principles are realized in practice. With core values such as human-centricity, transparency, and accountability at its heart, the framework aims to produce collaborative intelligence systems that not only perform effectively but also deserve the trust of those who depend on them. The next section will discuss how organizations can implement this framework concretely, integrating it into existing processes and governance structures.

5. Framework Implementation

Implementing Trust by Design in real-world projects requires integrating its principles and processes into the workflows of system design, development, and deployment. In line with our third research question, this section outlines how organizations and engineering teams can operationalize the framework: from embedding it in design and development processes and conducting ethical risk assessments to establishing governance and responsibility and measuring ethical performance. A comprehensive implementation checklist is provided in Appendix A to guide organizations through this integration process.

5.1. Integration into Design and Development Processes

Adopting Trust by Design starts with treating ethical and trust requirements as first-class requirements alongside functional and performance requirements. Teams should begin every collaborative intelligence project by explicitly identifying ethical goals (e.g., "ensure the robot's actions are interpretable by users" or "the AI's recommendations must be fair between team members"). These can be captured in requirement documents or user stories. Modern development methodologies such as Agile or DevOps can incorporate ethics checkpoints in their cycles. For instance, during each sprint review, the team assesses not only feature completion but also whether the implementation meets the trust-by-

design criteria—perhaps using a checklist derived from the core principles (transparency, privacy, etc.). Design documents would include sections addressing how the system design addresses each ethical principle (similar to how safety-critical systems include safety cases). For a structured approach to identifying and tracking these ethical requirements, refer to the Initial Setup and Planning section in Appendix A.

One practical tool is an Ethical Design Canvas, Figure 6, such as the one presented by [67], or similar frameworks, where designers map out stakeholders, possible harms, and mitigation strategies at the ideation phase. Another is user journey mapping [68], which includes emotional and trust-related states—mapping how a worker might initially be wary of an AI system and what features or support can move them to confidence. By making these considerations visual and explicit, the team keeps ethics in scope throughout development.

Figure 6. Dimensions of an Ethical Design Canvas.

During implementation, cross-functional teams are beneficial—including not just engineers, but also HR (for worker perspectives), safety officers, and ethicists or legal experts if available. This ensures that diverse aspects (from psychological safety to data compliance) are considered in design trade-offs. For example, a development decision about logging detailed user data for analytics might be checked by a privacy officer for necessity and compliance with regulations. Embedding such multi-disciplinary review in the dev process prevents issues from being discovered post-hoc.

Moreover, simulation and prototyping are used to test ethical aspects early. A VR simulation of a human–robot collaboration could reveal if the robot's movements are intimidating or if the user interface confuses the operator, as it allows iterative improvement before finalizing design [69]. Prototyping explanation UIs for AI decisions with real users can show which explanations actually increase understanding. The key is to iterate not just on technical performance, but on user trust outcomes.

5.2. Ethical Risk Assessment Methodologies

Similar to how projects conduct risk assessments for safety or business continuity, Trust by Design calls for ethical risk assessments. This is a systematic identification of potential ethical and trust failure modes. For collaborative systems, some risk examples could be "The AI might recommend an action that violates a safety procedure", "The robot might be misperceived as surveillance by workers", or "In case of network failure, the human loses crucial info and makes a poor decision". Each identified risk is analyzed for likelihood and impact. Techniques such as scenario analysis, what-if brainstorming, and even failure mode and effects analysis (FMEA) can be repurposed for ethical dimensions [70], which could be referred to as an "Ethical FMEA".

For each risk, mitigations are devised. If the risk is AI recommending unsafe actions, a mitigation could be implementing rule-based safety checks (do not allow recommendations beyond certain thresholds) and requiring human confirmation for high-impact decisions. If the risk is worker misperception of surveillance, the mitigation might be to clearly communicate what data the cobot does and does not capture and perhaps include a physical indicator (such as an LED) when it is recording data, plus giving workers control to pause data collection during breaks. Each mitigation is then implemented or documented as a limitation if it cannot be fully resolved.

This process should align with existing safety risk assessments. In fact, combining them might be efficient—consider a unified "Ethical and Social Impact Assessment" document that covers privacy, bias, and psychosocial factors alongside traditional safety. UNESCO has advocated tools for AI ethical impact assessment that guide evaluating benefits and risks relative to values and principles [71]; organizations can adapt such guidelines to their internal processes.

Throughout the project, revisit the ethical risk register as new features are added or when deploying in new contexts. For example, if a collaborative system that was tested in one plant is to be scaled company-wide or to a different country, reassess because cultural differences or new worker groups might introduce new ethical risks (such as differing perceptions of automation). Appendix A provides a detailed checklist for conducting comprehensive ethical risk assessments, including identification, analysis, and mitigation steps.

5.3. Governance Structures and Responsibility Allocation

A clear governance framework is necessary to uphold Trust by Design principles organization wide. Companies should assign responsibility for ethical oversight of collaborative intelligence projects. This could be a designated ethics officer or an AI ethics committee that reviews projects at key milestones. Alternatively, some organizations create working groups that include management, worker representatives, and experts to oversee Industry 5.0 implementations.

Governance also means defining roles: Who is responsible for monitoring the AI's outputs daily? (e.g., a shift manager) Who owns the data and ensures it is managed properly (perhaps a data steward)? Who should employees contact if they have concerns about the AI/robot's behavior? (maybe an ombudsperson or the ethics officer). By allocating such responsibilities, the organization signals that it takes these issues seriously and has mechanisms to address them, which in turn fosters trust among employees.

A multi-stakeholder governance approach is recommended; this means involving different levels of the organization and even external voices (such as domain experts or ethicists) in policy-making for AI use [72,73]. For example, a governance policy might stipulate that any introduction of collaborative robots must involve consultation with worker unions or safety committees and that the deployment plan must be approved by the ethics committee. This ensures broad buy-in and that no single viewpoint dominates the decision (which could overlook important concerns).

Regular governance meetings (quarterly, annually) can review the ethical performance metrics (discussed below) and any incidents or near misses. If a pattern of minor issues is noticed (say multiple instances of users overriding the AI due to distrust), the governance body can mandate a deeper investigation or adjustments to the system/training. A structured approach to establishing governance roles and procedures is outlined in the Governance and Responsibility section of Appendix A.

5.4. Metrics for Measuring Ethical Performance

What gets measured gets managed. To know if Trust by Design is effective, organizations should track certain KPIs (Key Performance Indicators) related to ethics and trust. Possible metrics, illustrated in Figure 7, include:

- User trust levels: Measured via periodic surveys or interviews. Questions can gauge confidence in the system, perceived transparency, perceived impact on job satisfaction, etc. For instance, a trust index might be compiled from statements such as "I can predict how the robot will behave" or "The AI's recommendations are generally sensible" rated by users. High trust scores (with healthy calibration—not overtrust) indicate success.
- Usage and override statistics: How often do users follow the AI recommendations vs. override them? How frequently do they resort to manual control of a cobot? If overrides are extremely high, it might indicate a lack of trust or usefulness. If overrides are zero but there were some AI errors that went unchecked, it could indicate over-trust or complacency. Balanced behavior where users appropriately rely on the system most of the time but occasionally correct it when needed would show well-calibrated trust.
- Incidents and near misses: Track any safety incidents or ethical issues (like a time the AI made a biased suggestion that was caught). Even if no actual harm occurred, near-miss reporting is invaluable. A log of "the AI almost caused X, but a human caught it" or "a worker felt uncomfortable with Y scenario" helps identify weak points. The goal is to see these numbers trend down as the system and training improve. An increasing trend would signal a need for intervention.
- Complaints or feedback tickets: If the company has a channel for employees to express concerns about technology, the number and nature of complaints related to the collaborative system is a metric. For example, if privacy complaints drop to zero after an update that clarified data use, that's a win.
- Diversity and fairness metrics: Analyze system outcomes for potential bias. For example, if it is an AI allocating shifts or maintenance tasks, measure distribution across employees to see if any group is overburdened. If it is a quality control AI flagging human work, ensure no particular worker's outputs are flagged disproportionately without explanation. Fairness metrics could include statistical parity indices or disparate impact ratios drawn from the AI ethics literature, applied to the specific context.
- Adoption and retention: Indirectly, trust is reflected in continued usage. Metrics such as how many tasks are successfully handled by human-AI teams versus reverted to manual processes can indicate acceptance. In training contexts, whether new employees are quick to learn the system can reflect its intuitiveness (a proxy for good design). Even employee retention or attrition rates in teams using the new system versus those that do not, could be insightful—ideally, the introduction of collaborative intelligence does not drive people to quit and perhaps even improves retention if it makes jobs easier or more engaging.
- Performance with ethical constraints: If the system uses multi-objective optimization including ethical factors, measure how well it is balancing them. For example, a scheduling system might have a target of no worker getting more than X hours of strenuous work. The metric would be the percentage of schedules adhering to that. Meeting ethical targets while achieving business goals demonstrates the framework's success.

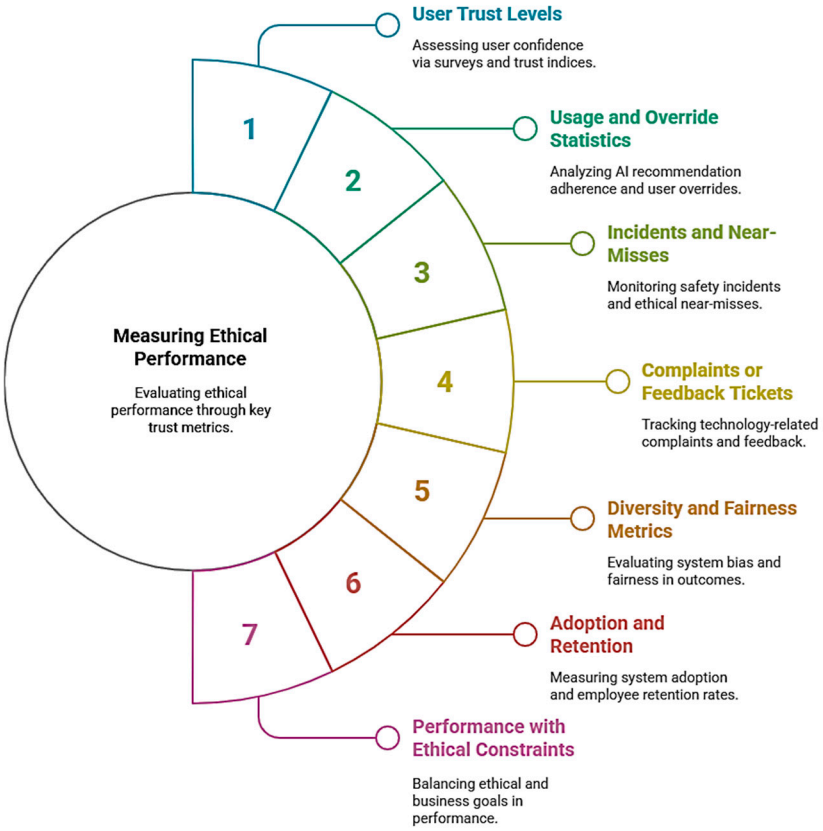

Figure 7. Key metrics for measuring ethical performance in Trust by Design implementation.

All these metrics should be reviewed in management meetings. It may help to create a dashboard that anonymizes and aggregates relevant data for decision-makers and possibly for employees too (transparency in metrics can further build trust—e.g., sharing the statistic that "100% of AI-driven decisions this month were reviewed by a human" or "no safety incidents in 2000 h of human-robot collaboration" gives confidence to all stakeholders). When negative metrics appear, the organization should respond proactively—for instance, if surveys show lower trust in a particular department, engage with those workers to understand why (maybe they had a specific bad experience) and address it via system tweaks or additional training.

Importantly, these metrics feed back into the continuous improvement loop. Trust by Design is not static; if metrics show areas for improvement, the framework dictates that the team revisit the design or processes to enhance trust and ethics. In this way, ethical performance management becomes part of normal operational excellence programs.

As to the adaptability of the framework to cater across sectors and organizational cultures. The framework is intentionally sector-agnostic but not sector-blind. High-reliability industries (aviation, healthcare) typically favor hard safeguards—pre-deployment certification, mandatory human-override tiers—whereas light-manufacturing or creative sectors lean on soft governance such as peer review and agile iteration. Cultural context also matters: in collectivist settings, shared accountability boards gain traction, whereas individualist workplaces prioritize personal override controls [74]. Table 3 maps these contingencies to the trust-by-design lifecycle, demonstrating parameter "dials" (e.g., audit frequency, explanation depth) that practitioners can tune rather than reinventing the entire framework.

Table 3. Contextual calibration of trust-by-design "dials" across lifecycle stages.

Lifecycle Stage	Dial	High-Reliability Industries	Light-Manufacturing/Creative Sectors	Collectivist Cultures	Individualist Cultures
Design and Development	Certification rigor	Mandatory, formal certification	Optional, guideline-based	Approval by shared committee	Individual sign-off with oversight
	Stakeholder involvement	Broad, cross-functional review boards	Lean, agile prototyping workshops	Group-focused co-design sessions	Empowered experts driving design
Testing and Validation	Audit frequency	Quarterly independent audits	Ad hoc peer reviews	Rotating team review rotations	One-on-one expert debriefs
	Testing depth	Comprehensive in-situ trials	Minimal viable testing	Consensus-based pilot groups	Self-directed sandbox testing
Deployment	Override mandate	Mandatory human-override layers	Optional "on-demand" override buttons	Shared decision-making boards	Personal override controls
	Onboarding training	Formal certification courses	Informal workshops and demos	Group training sessions	Self-paced e-learning modules
Operation and Maintenance	Monitoring intensity	Continuous real-time monitoring	Periodic spot checks	Team monitoring rosters	Personal dashboards and alerts
	Transparency level	Detailed logs and dashboards	High-level summaries	Collective reporting sessions	Individual notifications
Feedback and Evolution	Feedback loop formalization	Scheduled, structured review cycles	Open, rolling feedback channels	Committee-driven retrospectives	Direct feedback to system owners
	Adjustment cycle	Fixed quarterly updates	Continuous integration/deployment	Collective roadmap planning	Individual-driven feature requests

It is important to bear in mind that Trust by Design is not a one-size-fits-all recipe. Table 3 outlines how organizations can 'dial' elements such as audit rigor, override mandates, and training depth to match industry reliability requirements and cultural contexts. This built-in flexibility allows the framework to be tailored rather than universally imposed.

By integrating Trust by Design into everyday processes, conducting thorough ethical risk assessments, setting up governance with clear accountability, and measuring outcomes, organizations can institutionalize ethical collaboration. It transforms ethics from abstract principles into concrete practices and accountability structures. The next section aims to demonstrate how these implementations can work in practice by examining representative case studies and scenarios where human-machine collaboration is deployed, highlighting how the Trust by Design framework would handle real-world dilemmas and decisions.

6. Application Scenarios

To illustrate the Trust by Design framework in action, we present three vignette-based Industry 5.0 scenarios across different domains. These highlight recurring ethical dilemmas in human-machine collaboration and how our proposed framework can address them. Rather than focusing on one specific sector, we generalize lessons applicable across manufacturing, industrial decision support, and human augmentation contexts.

6.1. Scenario 1: Collaborative Robots in Manufacturing

An automotive factory deploys collaborative robots (cobots) on the assembly line to work alongside humans in installing heavy components. The cobots can intelligently hand tools to workers, hold parts in place, or perform repetitive torquing operations. Workers initially have concerns about safety (working in close proximity to moving robots) and job security (wondering if these cobots will gradually take over their tasks). The company's

goal is to improve production efficiency and reduce ergonomic injuries for workers while maintaining a human-centered environment.

Ethical challenges and trust issues: In this scenario, physical safety is paramount. Even though cobots are designed with safety features (force-limited joints, sensors to stop if a human is too close), workers may not immediately trust that the robot will stop in time to avoid a collision. There's also role ambiguity—if the cobot can do part of a task automatically, workers might be unsure when to intervene, potentially leading to either overreliance or under-utilization of the robot. Job displacement fears are present, since the workers see robots doing tasks they used to do manually. Additionally, accountability questions arise: if a cobot-installed part later fails quality inspection, is the worker or the automation at fault? All these factors can impact trust and morale.

Trust by design application: The company applies our framework from the outset. During system design, engineers and safety managers, together with worker representatives, perform an ethical risk assessment. For safety, they identify risks such as "cobot accidentally hitting a worker's hand" and ensure mitigations: the cobot's speed is capped when near humans, and it is programmed to maintain a minimum distance unless explicitly in cooperative mode. They also implement a simple intention signaling system on the cobot—e.g., a light or small display that indicates its next action ("Moving door panel into position") so the worker is never surprised by its movements, addressing transparency in real-time. Early prototypes are shown to some veteran workers, who give feedback that they would like a manual override pendant they can carry. The designers incorporate this: each worker has a portable emergency stop or pause button for the nearest cobot, giving them a sense of control (human agency principle).

Before full deployment, a training program is conducted. Instead of just technical training, it also covers psychological aspects: instructors explain that the cobots are there to assist and not replace; they show how certain injury-prone tasks will now be handled by cobots (such as overhead drilling) and emphasize that this allows workers to focus on finer assembly and quality checks. They even share data from pilots indicating reduced worker fatigue and maintained production output to build a narrative that cobots are enabling a better workplace, not threatening it (aligning with the human-centric value). By being transparent about objectives and outcomes, management builds trust.

During the pilot phase at one assembly station, feedback mechanisms are in place. Workers can report any uneasy incidents or suggestions via a tablet stationed nearby. One early observation is that when two workers and one cobot collaborate on a task, the workers sometimes are not sure if the other person or the cobot will perform the next step, causing brief confusion. This is a coordination issue—the team refines the standard operating procedure and perhaps adds an audio cue from the cobot at certain handover points (such as a subtle chime when it is completed with its action and expects the human to take over). This kind of fine-tuning exemplifies adjusting the interface of collaboration to maintain clarity and trust.

From a stakeholder perspective, to address job security fears, management could make a commitment (possibly in agreement with unions) that no layoffs will result from cobot introduction, and instead any productivity gains will be used to improve work conditions or upskill workers for higher responsibilities. Keeping this promise is vital—trust in technology is intertwined with trust in the organization. When workers see that the cobots genuinely reduce their physical strain and that they are still valued (perhaps their roles evolve to supervising multiple cobots or doing more complex craftsmanship), trust grows. Over time, workers might start to trust the cobots like team members—e.g., knowing that "the robot will always hold this part steady for me, so I can focus on bolt tightening".

In terms of accountability and responsiveness, suppose a minor incident occurs—say a cobot brushes a worker because a sensor was briefly obstructed. The framework's governance kicks in, the incident is logged, investigated transparently, and revealed to be low impact but nonetheless used as a learning moment. The company communicates openly with the team about what happened and updates the cobot's sensor cleaning schedule to prevent reoccurrence. This response would further reinforce trust: workers see that the system and managers are accountable and committed to safety.

In summary, the manufacturing case shows how Trust by Design handles safety ethics (through design and training), transparency (through signals and procedures), autonomy (keeping human override control), and psychosocial factors (through communication and policy) to make human-cobot collaboration successful. The outcome is a resilient human–robot team where humans trust the robots to do their part safely and effectively, and robots effectively augment human labor without diminishing human agency or value.

6.2. Scenario 2: AI Decision Support in Industrial Operations

A chemical processing plant introduces an AI-driven decision support system to assist control room operators in managing complex processes. The AI analyzes sensor data (pressure, temperature, flows) and suggests adjustments to optimize yield and prevent faults. It can predict anomalies hours in advance. Operators remain in charge but are expected to consult the AI's recommendations for routine and emergency decisions. This is a high-stakes environment—wrong decisions can cause equipment damage or safety incidents. The challenge is to ensure operators trust and correctly utilize the AI without becoming over-reliant on it and that the AI's advice is ethically and technically sound.

Ethical challenges and trust issues: Here, transparency of reasoning is a major challenge—operators with decades of experience may find it hard to trust a "black-box" AI telling them to, say, lower a reactor temperature preemptively. If the AI cannot explain its prediction in ways that align with the operator's mental model, they might ignore potentially life-saving advice (under-trust). Conversely, if they come to over-trust the AI, they might follow a bad recommendation without double-checking, possibly leading to an accident (over-trust). Ensuring the right balance of control is tricky: the company does not want to remove the human from the loop but also does not want the AI's benefits wasted due to mistrust. Accountability is also complex: if an operator follows AI advice that leads to a bad outcome, who is responsible? The operator might blame the AI, but ultimately the company is accountable for decisions. There's a risk of the moral crumple zone effect—the operator might be scapegoated because the AI cannot be "blamed," which would be ethically problematic. Privacy is less of an issue with machine data, but safety and reliability of the AI are critical ethical imperatives (the AI must be rigorously validated to not inadvertently suggest unsafe actions).

Trust by design application: Following the framework, the AI system is to be developed with extensive input from the operators (end-users) and process engineers. Early on, the designers adopt an explainable AI model—for instance, instead of a pure black-box neural network, they use a hybrid model that can highlight which sensor readings or trends most influenced its suggestion. In the UI, when the AI suggests an action, it accompanies it with a rationale like "Reactor pressure trending high and catalyst aging detected; recommend reducing feed rate by 5%". The system might even show a graph comparing current trends to historical incident patterns to justify the recommendation. This aligns with the transparency principle, giving operators insight into the AI's reasoning.

To maintain human agency, the system is to be configured as "advisor", not "autopilot". It cannot directly control actuators; it presents advice that the human must approve and implement. The interface is designed to make the human the ultimate decision-maker:

for each recommendation, the operator can choose "Accept", "Modify", or "Reject" and must confirm the action. This keeps them actively engaged and avoids blind automation. However, to guard against potential automation complacency (operators getting lazy due to AI always being right), the training program must include scenarios where the AI makes a suboptimal suggestion and the trainees learn to recognize it and override—reinforcing that human judgment is still critical. Over time, this training in a simulator builds trust: operators see that the AI is usually right but also understand its limits and how to double-check critical suggestions.

To ensure reliability, the AI would have to undergo extensive testing with historical data, being shadow tested live (giving recommendations that were observed but not enacted) to verify that it rarely conflicts with expert human judgment except when it genuinely catches something humans overlooked. During deployment, ethical performance metrics have to be monitored: how often do operators agree with the AI? Are there cases where the AI was right and the human ignored it, and why? If an operator consistently rejects advice, supervisors engage to see if there's a trust issue or model issue. Perhaps the operator notices the AI does not account for a certain nuanced condition—this feedback can be used to refine the AI (continuous improvement). Conversely, if operators start rubber-stamping all AI suggestions without analysis, more training or interface tweaks might be needed to encourage thoughtful review (maybe by occasionally requiring a reason for accepting in critical situations to ensure they considered it).

One specific ethical design feature: the AI is constrained never to suggest actions outside of safety limits or standard operating bounds (a rule-based safety overlay). For example, it will never suggest opening a valve beyond the allowable limit or mixing chemicals in a ratio that violates regulations. This deontological safeguard ensures that even if the AI's optimization engine somehow thought an extreme action would optimize yield, it would not present that to the user. Thus, the human is never put in a position of considering an unethical or unsafe recommendation from the AI—maintaining trust that "the AI will not lead me astray from fundamental safety rules".

6.3. Scenario 3: Human Augmentation Technologies in Industrial Settings

A logistics company equips its warehouse workers with wearable exoskeletons (powered suits that support the back and arms) and AR (augmented reality) smart glasses. The exoskeletons reduce strain when lifting heavy items, and the AR glasses provide real-time information such as item locations and optimal stacking patterns. These technologies represent human augmentation aimed at improving productivity and safety. Ethical questions arise about mandating use, potential health effects, and privacy (the AR glasses have cameras scanning the environment, which could be seen as surveillance). For this scenario, we will focus on the exoskeleton aspect for a specific ethical dilemma: Should wearing the exoskeleton be compulsory for certain tasks, and how should we handle workers who find it uncomfortable?

Ethical challenges and trust issues: Exoskeletons blur the line between human and machine. A key issue is autonomy and consent: some workers may not want to wear the device due to discomfort, pride in manual ability, or distrust of new tech. If the employer mandates it to prevent injuries, is that ethically acceptable? Pote et. al. (2023) in [75] noted that forcing workers to wear an exoskeleton like a piece of PPE raises concerns if it does not fit well or causes pain—analogous to mandating an ill-fitting safety jacket that causes bruising. Ethically, the exoskeleton should truly help and not be an instrument of exploitation enabling the company to push workers to lift even more weight. There's a trust component: workers need to trust that the device is safe (would not injure them or malfunction). If early versions are clunky or cause any pain, trust in the technology

will plummet. Also, data privacy might come in if the exoskeleton or glasses collect performance data—workers might fear it is used for monitoring their speed or movements for disciplinary purposes, not just assistance. The AR glasses with cameras may feel like a surveillance tool as well, potentially eroding trust in management's intentions. Lastly, there's fairness: if some workers cannot use the exoskeleton (due to body shape or health contraindications), will they be disadvantaged in assignments or expected to do the heavy work without support?

Trust by design application: The company would roll out the augmentation tech with a voluntary pilot program first, rather than immediate mandatory use. This respects autonomy and allows the tech to earn worker buy-in. In the pilot, 10 workers try the exoskeleton and AR glasses. Their feedback is actively solicited: How does it feel? Did it make tasks easier? Any pain points? Suppose some say the exoskeleton shoulder straps dig in after a while. Engineers work with the exoskeleton vendor to adjust the fit or padding (very much a user-centered design fix). This iterative improvement is crucial so that by the time of broader deployment, most ergonomic issues are resolved, showing workers that their comfort is a top priority—an important trust signal.

The company also addresses policy transparently: they tell workers that the exoskeletons are intended to reduce injuries and not to increase workload weight limits without medical evaluation. In fact, they might put in writing that maximum box weight limits will remain the same or even be lowered because the exoskeletons provide an additional margin of safety. This counters any perception that the technology will be used to squeeze more labor out of them. It aligns with the principle of beneficence (we're doing this for your well-being) and justice (not using tech to impose unfair demands).

For the AR glasses and exos, they implement privacy safeguards—e.g., the AR system processes visual data on the device itself to guide the worker but does not stream video to management. Any performance data collected (such as number of lifts) is shared with the worker themselves and used in aggregate for process improvement, not for individual monitoring for punishment. These rules are codified in an agreement. By making these guarantees, management builds trust that "these tools are here to help you, not spy on you".

They also follow an inclusive approach: if some workers cannot use the exoskeleton (due to medical implants or it does not fit), they ensure those workers either get alternative accommodations or are not penalized. Perhaps they can use other assistive devices or have team lifts. Fairness is maintained by, for example, rotating tasks so that those using exoskeletons are not always given all the heavy lifts—everyone still shares the workload in a reasonable way, with exos augmenting everyone as needed. The device is presented as a benefit or optional support—at least initially. Over time, if the majority find it beneficial, social proof and slightly improved productivity might naturally make usage widespread without top-down mandates.

One ethical dilemma was whether to make it mandatory. Suppose over months, data shows back injuries have dropped among those using exoskeletons, and those not using them are getting more strains. The company might consider requiring it for certain high-risk tasks. Trust by Design would approach this carefully: engage with workers and possibly health and safety reps to discuss the findings. If the evidence strongly favors use, a consensus might emerge that "for your own safety, we should all use them when doing X task". They could then implement a policy that whenever lifting above Y kg, the exoskeleton must be worn, similar to how certain PPE is required. The difference here is they reached that decision collaboratively with clear evidence, rather than arbitrarily imposing it. They also keep monitoring the comfort and health impacts of long-term use (e.g., does wearing it all day cause fatigue or any unintended issues?).

Meanwhile, the Trust by Design engineering of the exoskeleton itself might include alarms if the user is moving in a way that could cause injury despite the exos—such as bending incorrectly—effectively coaching the user (e.g., a gentle vibration reminding them to lift with the device's support). This kind of feature is ethically tricky (could feel like the device is nagging), so it is implemented only if users want it and find it helpful. Ideally, it is customizable (the user can turn off the coaching if they find it annoying). Giving users control over such features respects their autonomy and encourages trust in the device—it is a tool serving them, not controlling them.

Outcome: Over time, most warehouse workers adopt the exoskeletons as a standard part of their gear because they feel the difference: less fatigue, fewer aches. They trust that the device is making their job safer. New hires see the positive attitude of veterans and quickly accept the technology as well. The few who initially resisted either come around after seeing colleagues' benefit, or if someone still cannot use it, the company ensures that person is not disadvantaged or perhaps moves them to a role with less heavy lifting (with no loss of pay—showing the company values the employee's health over forcing tech). Because the rollout was performed with respect and involvement, there was not an adversarial dynamic. One can imagine a counterfactual where, if management had just dumped the gear and said, "you must wear this or face discipline", workers might distrust the equipment, wear it improperly, or try to game it, defeating the purpose. Trust by Design avoided that by building a positive feedback loop of trust—employees saw management genuinely cared (taking feedback, not using data against them), and management saw employees engaging constructively with the tech, not sabotaging or ignoring it.

Across these scenarios, a common thread is that ethics and trust go hand in hand. By proactively addressing ethical issues (safety, transparency, fairness, and autonomy), the organizations built systems and policies that employees and operators could trust, which in turn led to better adoption and outcomes. The Trust by Design framework provided a structured approach to foresee and manage ethical dilemmas, whether it is deciding how an AI should explain itself, how a robot signals intentions, or how to set policies for wearable tech.

In all cases, when a potential ethical dilemma was considered in the projection of the scenarios, the framework suggested inclusive dialogue and empirical evaluation rather than top-down enforcement. This not only tends to yield ethically sound decisions but also aims to strengthen trust by making people feel that their perspectives are heard and the resulting solutions are fair and sensible. The next section will zoom out to the regulatory and policy landscape to place these organizational practices in context with broader legal and standardization efforts shaping Industry 5.0 and collaborative intelligence.

7. Regulatory and Policy Implications

The rise of collaborative intelligence in Industry 5.0 has prompted regulators and policymakers to consider how existing laws apply and where new rules or standards are needed. Ensuring that human-centric and trustworthy practices are followed at scale may require more than voluntary frameworks; it might demand formal regulations, especially in matters of safety, privacy, and labor rights. In this section, we examine the current regulatory landscape in major jurisdictions, offer recommendations for policy development, discuss emerging industry standards, and consider the importance of international harmonization.

7.1. Comparative Regulatory Approaches in Europe and the United States

The European Union champions an ethical, human-centric Industry 5.0 agenda. Beyond the broad 2021 policy brief [76], the forthcoming AI Act [77]—phased in from 2026—will be the first law to tier AI by risk. Industrial worker-management tools land

in the "high-risk" tier, triggering strict rules on transparency, traceability, accuracy, and mandatory human oversight (Art. 14). GDPR adds further limits whenever personal data—say, wearables or cameras—is used, while long-standing safety rules (e.g., the Machinery Directive) already require CE-marked compliance for cobots or exoskeletons. An updated Machinery Regulation will soon extend that duty to AI-enabled equipment. Overall, Europe is turning principles—oversight, transparency, and non-discrimination—into hard obligations: firms must assess risks, document human control, and often certify or register their AI. Simultaneously, EU funds, digital-innovation hubs, and standards initiatives incentivize the shift to human-centric manufacturing.

Conversely, the United States oversight remains patchwork and sector-specific. Instead of a single AI statute, federal guidance leans on voluntary tools—e.g., NIST's 2023 AI Risk Management Framework [78]—and post-hoc enforcement. OSHA can cite employers under the General Duty Clause if a cobot or exoskeleton harms a worker, yet no pre-market robot rules exist beyond ANSI/RIA standards. Anti-bias watchdogs (EEOC, FTC) warn that algorithmic hiring or workforce tools must still satisfy Title VII and the FTC Act, while state privacy laws (e.g., California's CPRA) curb workplace monitoring. The DoD has adopted binding ethical-AI rules for contractors, and the White House's 2022 Blueprint for an AI Bill of Rights [79] urges agencies to safeguard safety, transparency, and fairness. Expect continued "soft-law" pressure—and the prospect of future regulation—around Industry 5.0 deployments.

In sum, Europe is moving toward legally enforcing trustworthiness in AI and Industry 5.0 systems, whereas the U.S. is using a combination of existing laws and new guidelines to similar effect. In both jurisdictions, accountability and human oversight of collaborative systems are central themes—matching the core of Trust by Design.

7.2. Industry Standards and Best Practices

Beyond law, industry groups and standardization bodies are publishing guidelines that effectively shape how Trust by Design can be implemented uniformly:

- The ISO and IEEE are notable: IEEE's 7000-series ethics standards—for example, IEEE 7001-2021 Transparency of Autonomous Systems and IEEE 7007-2021 Ontological Standard for Ethically Driven Robotics and Automation Systems—give engineers concrete, testable requirements for embedding ethical attributes [80,81]. ISO is working on an AI management system standard for AI governance, the ISO/IEC 42001:2023 Artificial Intelligence—Management System [82]. Companies can voluntarily adopt these to demonstrate their commitment to best practices. We recommend that industries adopt a certification approach—for example, a "Trustworthy AI" or "Collaborative System Safety" certification from a recognized body would signal to stakeholders (including insurers, clients, and employees) that the system meets a high ethical standard.
- Best practices sharing: Organizations such as the Robotics Industries Association (RIA) in the US or the International Federation of Robotics (IFR) often publish technical reports and case studies. For instance, guidance on implementing cobots safely or lessons learned from human-AI teamwork. Policymakers can encourage industry consortia to develop open guidelines—analogous to how the automotive industry shares safety test protocols. In the context of Industry 5.0, best practices might include how to involve employees in tech deployments or how to run an effective pilot. Companies should not have to reinvent the wheel ethically; documenting and sharing what works (such as effective training methods or interface designs that improved trust) can accelerate widespread adoption of trust-centric design.
- Another best practice is aligning corporate governance with these ideals: e.g., companies could incorporate ethical AI use into their ESG (Environmental, Social, Gov-

ernance) reporting. Already, some companies report diversity and safety metrics; adding AI ethics metrics (such as the number of AI systems assessed for bias or having ethics committees) could become part of social responsibility indices. This pressures companies to follow frameworks such as Trust by Design to meet investor and public expectations.

International harmonization considerations: Industry 5.0 is a global movement; manufacturers and tech providers operate across borders. Disparate regulations can both hinder innovation and reduce clarity on ethical obligations. It is in everyone's interest to strive for harmonized standards so that a system considered trustworthy in one country is not deemed unethical in another. The OECD's AI Principles (backed by 40+ countries) provide a high-level consensus on values such as human-centeredness and robustness. These have informed both EU and US policies and could serve as a basis for aligning efforts.

In conclusion, regulation and policy are catching up to the rapid technological advances with a clear trend: embedding trust and ethics as requirements, not mere options. Organizations practicing Trust by Design will likely find themselves well-positioned to comply with emerging laws (since they've proactively addressed human oversight, transparency, etc.), whereas those who ignore ethics may face compliance headaches or liabilities down the line. Our recommendation is for a coordinated approach: policymakers set the guardrails and incentives for ethical practice, industry standards provide the technical playbook, and companies implement these on the ground—all informed by continual dialogue with the workforce and public. This multi-layered governance ensures that collaborative intelligence systems truly earn the label of "trustworthy AI" across the world.

Finally, as both regulation and technology evolve, there will be new challenges and questions. The next section looks ahead to future research directions, anticipating the ethical issues of tomorrow's Industry 5.0 and how our framework might adapt.

8. Future Research Directions

Industry 5.0 and collaborative intelligence are still emerging fields, and with rapid technological innovation come new ethical challenges and unknowns. To keep the Trust by Design framework relevant and robust, ongoing research and cross-disciplinary collaboration will be necessary. In this section, we outline several future research directions: anticipating emerging ethical challenges, considering technological developments that may impact the framework, exploring cross-disciplinary research opportunities, and discussing how the framework might evolve and adapt.

8.1. Emerging Ethical Challenges

As collaborative intelligence systems become more advanced, new ethical issues will surface. One area is the impact of AI decision support on human expertise over time. For example, if AI advisors handle routine decisions, human operators might lose skill or situational awareness—a phenomenon known as "skill fade" or automation complacency [83]. Ethically, how do we ensure humans remain capable of taking over when needed? Research could explore training regimes or AI designs that deliberately keep humans in the loop enough to maintain expertise (perhaps by occasionally deferring decisions to humans even when not strictly necessary). Another emerging issue is the emotional and social effects of working with AI and robots. Early studies in human–robot interaction show people can develop emotional attachments to robots or treat AI with anthropomorphic qualities. What are the implications of workers developing trust or friendship with a robot colleague? Could that be leveraged positively (to encourage safety compliance, for instance), or might it be manipulative to design robots to elicit emotions (raising concerns of deception)? Future

research can empirically study these dynamics, guiding ethical design to foster appropriate social relationships with machines (neither cold mistrust nor unhealthy attachment).

The increasing integration of biometric [84] and AI-driven health analytics in the workplace also poses challenges—imagine sensors that detect stress or fatigue [85] and AI that suggests a break or task rotation. While beneficial, this treads on personal data. Studies on how to do this in a worker-respecting way (perhaps by keeping data on the device or giving control to the employee) would be valuable. Another frontier is AI-driven adaptive management—e.g., algorithms scheduling work dynamically. There's an ethical imperative to ensure these algorithms are fair and do not inadvertently exploit workers. Ongoing research is needed as to how workers perceive algorithmic management and what aspects are most problematic (lack of explanation? inability to negotiate?). Some work has begun in the gig economy context, but as Industry 5.0 brings AI management into factories and warehouses, this research needs to expand.

8.2. Technological Developments Impacting the Framework

Emerging tech in AI and robotics will test and extend Trust by Design. For instance, more generalized AI or autonomous systems might take on more complex roles, making it harder to predict all possible behaviors or decisions (thus challenging transparency and safety assurances). We may need new methods for verifying and validating such systems—perhaps simulation-based ethical stress testing, where AI is put through thousands of simulated scenarios, including edge cases, to see how it behaves. If quantum computing or more powerful AI allows analyzing systems in real-time for near-optimal decisions, human roles might shift to oversight of multiple processes—raising questions of span of control and cognitive load. Technological research on better human-AI interfaces (such as conversational AI that operators can question) will be key to maintaining trust when the AI's reasoning complexity far exceeds human cognitive capacity.

8.3. Cross-Disciplinary Research Opportunities

Addressing these complex questions will require collaboration between engineering, psychology, ethics, law, and sociology. For example, understanding trust deeply might involve psychologists and neuroscientists studying how humans build trust with non-human agents and what design features in AI evoke trusting behaviors without false overconfidence [7]. Ethicists and legal scholars can help translate moral principles into operational criteria that engineers can implement (such as defining what counts as an explanation that satisfies a duty of transparency). Sociologists and labor economists might study the broader impacts on work culture and job quality, which in turn informs what outcomes we consider ethical success (such as does collaborative tech lead to higher job satisfaction? If not, why, and how to improve it?).

One concrete cross-disciplinary research idea could consider the development of a "Trustworthiness Index" for collaborative systems, combining technical measures (reliability, safety), user perceptions (survey scores), and outcomes (accident rates, productivity changes). This could be akin to a Consumer Reports or UL safety rating but for trust. It would need input from statisticians, social scientists, and engineers to create a valid and reliable measure. Having such an index would also allow longitudinal studies—e.g., comparing across companies or countries—and could motivate improvements.

Another area is participatory design research involving ethicists and workers in co-creating new tools. For instance, running living labs in actual factories where new AI or cobots are introduced and studied in real conditions, with ethicists on the team recording how values are negotiated on the ground, could yield rich insights beyond

what lab experiments offer. This would help refine frameworks similar to ours with real-world evidence.

8.4. Framework Evolution and Adaptation

The Trust by Design framework itself must remain flexible. It should be updated as norms and expectations evolve. For example, what is considered a sufficient explanation today might not satisfy tomorrow's more AI-literate workforce, so the standards for transparency may rise. If future generations grow up interacting with AI from childhood, their trust calibration could differ, requiring adaptation in how systems communicate or how much autonomy they are given. The framework might also extend to new domains: Industry 5.0 could expand beyond manufacturing to service industries (collaborative AI in healthcare, and education). Core principles would remain, but their implementation could look different (for instance, trust with a medical diagnostic AI involves patients and doctors both—adding another stakeholder layer).

Periodic reviews of the framework, possibly by an interdisciplinary panel, could identify gaps. Perhaps in five years, issues such as the environmental impact of AI (energy usage of running all these systems) will become a pressing ethical concern linking Industry 5.0 to sustainability goals. Then the framework might explicitly include an environmental stewardship principle, ensuring that collaborative systems are also green by design, since sustainability is one of Industry 5.0's pillars.

Another adaptative element is integrating with AI governance tools that may become standard. If companies start widely adopting AI audit tools or continuous monitoring platforms (some AI companies are developing "dashboard" products to track bias, drift, etc.), Trust by Design should incorporate those into practice—e.g., our implementation section would then advise deploying such tools and feeding their outputs into the ethical performance metrics.

In summary, research must continue to inform practice, and the framework should be considered a living guideline. By fostering close ties between researchers and practitioners (e.g., publishing results of case studies similar to ours or engaging in industry-academia partnerships to test new trust enhancement techniques), we ensure the framework does not stagnate. A virtuous cycle can be formed: field experience generates research questions, research produces new insights or technologies, which then update frameworks and standards, improving field practice further.

Ultimately, the vision of Industry 5.0 is a moving target—as technology and society change, the ethical framework guiding it must also progress. Investment in research and open sharing of lessons learned (including failures) will be crucial in the coming years. With a firm foundation and an agile approach to incorporating new knowledge, Trust by Design can evolve to meet future demands, ensuring that collaborative intelligence systems remain worthy of the trust we place in them and continue to serve humanity's best interests.

9. Limitations

This article delivers a conceptual trust-by-design framework grounded in a streamlined systematic review that deliberately incorporated academic papers, policy documents, industry white papers, and standards to give a broad, practice-oriented lens. Nevertheless, three boundaries remain:

Residual source bias. Although we went beyond traditional databases to include non-academic material, our search was restricted to English-language sources and to documents indexed in the chosen repositories. Important perspectives published in other languages or circulated only in niche region-specific venues may still be absent.

Temporal scope. The evidence base spans 2000–2025. Because ethical norms, best-practice guidelines, and regulations (e.g., the EU AI Act's implementation roadmap) continue to evolve, the framework will need periodic updates to remain aligned with current expectations.

Ecological validity. Scenario analysis cannot fully anticipate emergent human–AI behaviors, long-term trust calibration, or organizational dynamics that surface only during live deployments and extended use.

Trust breakdown scenarios. Though the framework embeds design and monitoring measures to promote trust, real-world deployments may still face persistent user resistance (under-trust), automation complacency (overreliance), performance degradation due to model drift (inconsistent reliability), or organizational barriers (insufficient training, siloed decision-making), which must be identified and addressed through empirical evaluation and adaptive governance.

Surveillance and power dynamics. While Trust by Design is intended to empower front-line users, its mechanisms (fine-grained audit logs, real-time dashboards, mandatory override records) could be repurposed by management as surveillance tools—tracking worker productivity, enforcing disciplinary measures, or reinforcing existing hierarchies under the guise of "safety" and "transparency". To guard against this misuse, we recommend:

(a) Role-based access controls on audit trails, so that only designated safety or ethics officers (not every manager) can view sensitive logs.
(b) Clear data use policies that prohibit performance monitoring or punitive use of trust metrics, with enforceable penalties for violation.
(c) Periodic "red-team" reviews by independent stakeholder representatives to ensure that data and controls remain aligned with user empowerment rather than managerial oversight.

Advancing from concept to evidence will involve pilot studies, longitudinal tracking of trust and override metrics, and cross-cultural expert appraisal. For now, the framework offers a coherent vocabulary and actionable design logic that researchers and practitioners can critique, adapt, and test, accelerating the collective move toward empirically grounded, ethically robust collaborative-intelligence systems.

10. Conclusions

Industry 5.0 represents more than a technological shift; it is a paradigm shift to align advanced collaborative systems with human values, sustainability, and resilience. In this paper, we have introduced Trust by Design, a lifecycle-driven approach that embeds ethical principles at every stage of human–machine collaboration. At its core, Trust by Design rests on human-centricity, transparency, privacy, autonomy, fairness, safety, and accountability, and translates these values into concrete practices: stakeholder co-design; ethical risk assessments; explainable, override-capable interfaces; rigorous safety and bias testing; and clear governance structures that assign responsibility and safeguard data use (Sections 3–5).

Through illustrative vignettes, from factory cobots reducing ergonomic injuries to AI decision support in control rooms and exoskeletons augmenting human strength, we have shown how these principles have the potential to foster calibrated trust, reduce both under- and overreliance, and promote user acceptance (Section 6). Crucially, the framework's contextual "dials" (Table 3) allow practitioners to tailor audit rigor, override mandates, training formats, and monitoring intensity to specific industry demands and cultural settings, ensuring the approach is flexible rather than prescriptive.

Moreover, the comparative review in Table 2 demonstrates that most existing studies address trust at particular levels or domains, such as ethics-by-design in software engineering, human-centric trust dynamics, or blockchain as a trust anchor, but do not span

the full lifecycle of collaborative intelligence systems. In contrast, Trust by Design unifies these different perspectives into a single, end-to-end framework and provides an actionable implementation checklist (Appendix A). While this checklist may not be groundbreaking in isolation, it offers a clear, practical bridge between high-level ethical concepts and everyday engineering practices, complementing, rather than competing with, prior approaches.

At the same time, we recognize the boundaries of our conceptual model. Its evidence base is drawn from English-language and indexed sources (2000–2025), and scenario analyses cannot capture every emergent behavior or organizational nuance. Real-world pilots, longitudinal trust metrics, and cross-cultural studies will be essential to validate and refine the framework (Section 9). Moreover, mechanisms designed for empowerment, such as fine-grained audit logs, must be governed to prevent misuse as surveillance tools.

Looking forward, Trust by Design is intended as a living guideline. As new technologies (e.g., more autonomous AI, quantum-powered analytics) and societal expectations evolve, the framework will need periodic updates, potentially adding principles such as environmental stewardship or integrating advanced AI-governance platforms (Section 5). By fostering ongoing dialogue among engineers, ethicists, policymakers, and workers, and by systematically measuring ethical performance, we can ensure that Industry 5.0 deployments are not only efficient but also deserving of the trust they require.

Ultimately, the success of Industry 5.0 will be measured not just in productivity metrics or ROI, but in the confidence and agency felt by those who work alongside these systems. Trust by Design offers a compass and a toolkit for action that aims to keep humans at the center, promoting adaptive resilience, and making trustworthiness a matter of design rather than expectation.

Author Contributions: Conceptualization, E.A.M.-C., I.G., M.S., R.G.R.-C., M.F.H. and S.S.; Formal analysis, E.A.M.-C., I.G., M.S., R.G.R.-C., M.F.H., S.S. and G.A.-C.; Investigation, M.S., M.F.H., S.S. and G.A.-C.; Methodology, E.A.M.-C., I.G., M.F.H. and R.G.R.-C.; Validation, R.G.R.-C. and G.A.-C.; Writing—original draft, E.A.M.-C., I.G. and R.G.R.-C.; Writing—review & editing, E.A.M.-C., I.G., M.F.H., S.S. and R.G.R.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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: The authors declare no conflicts of interest.

Appendix A. Trust by Design Implementation Checklist

This appendix presents a comprehensive checklist for implementing the Trust by Design framework. The checklist is designed to provide a practical tool for operationalizing ethical principles in the development and deployment of collaborative intelligence systems. It follows the lifecycle approach outlined in Section 4 of the paper, encompassing all stages from initial planning through continuous improvement.

The checklist is structured into nine key sections, each addressing critical components of the framework:

1. Initial Setup and Planning
2. Governance and Responsibility
3. Design and Development Stage
4. Ethical Risk Assessment
5. Testing and Validation Stage
6. Deployment Stage
7. Operation and Maintenance Stage

8. Feedback and Evolution Stage
9. Measurement and Reporting

Organizations can use this checklist as a reference guide when implementing collaborative intelligence systems that align with the Trust by Design principles. While not every item may be applicable to all contexts, the comprehensive nature of the checklist ensures that critical ethical considerations are not overlooked during implementation.

The items presented here derive directly from the theoretical framework outlined in the main text and represent a practical translation of abstract principles into actionable steps. By following this checklist, organizations can systematically embed trust-building mechanisms throughout the system lifecycle, ensuring that collaborative intelligence systems are developed and deployed in ways that inherently foster trustworthiness and ethical behavior.

Trust by Design Implementation Checklist

1. Initial Setup and Planning
   - ☐ Establish a cross-functional team including engineers, HR, safety officers, and ethics/legal experts
   - ☐ Define ethical goals and requirements alongside functional requirements
   - ☐ Create an Ethical Design Canvas mapping stakeholders, potential harms, and mitigation strategies
   - ☐ Develop user journey maps that include emotional and trust-related states
   - ☐ Establish project-specific Trust by Design metrics and KPIs
2. Governance and Responsibility
   - ☐ Designate ethics officer or form an AI ethics committee
   - ☐ Define clear roles and responsibilities for ethical oversight
   - ☐ Create a multi-stakeholder governance approach involving different organizational levels
   - ☐ Establish procedures for ethical review at key project milestones
   - ☐ Set up regular governance meetings (quarterly/annually) to review ethical performance
3. Design and Development Stage
   - ☐ Incorporate ethical user stories in requirements documentation
   - ☐ Perform simulations and modeling to foresee interaction patterns
   - ☐ Design redundant safety mechanisms
   - ☐ Integrate ethics checkpoints in development cycles (Agile/DevOps)
   - ☐ Create explanation panels or features for AI outputs
   - ☐ Design features that address each core principle:
   - ☐ Human agency (override functions and confirmation requests)
   - ☐ Transparency (explanations and confidence levels)
   - ☐ Privacy (data minimization, encryption, and access controls)
   - ☐ Fairness (bias mitigation techniques)
   - ☐ Safety (fail-safe mechanisms and predictable behavior)
   - ☐ Accountability (decision logs and audit trails)
   - ☐ User involvement (tutorials and training resources)
4. Ethical Risk Assessment
   - ☐ Conduct comprehensive ethical risk assessment or "Ethical FMEA"
   - ☐ Identify potential ethical and trust failure modes
   - ☐ Analyze each risk for likelihood and impact
   - ☐ Develop mitigations for each identified risk

- ☐ Document limitations for risks that cannot be fully resolved
- ☐ Align with existing safety risk assessments

5. Testing and Validation Stage
   - ☐ Conduct user testing specifically to gauge trust levels
   - ☐ Use trust scales or surveys to quantify user trust
   - ☐ Perform safety tests under various edge cases
   - ☐ Conduct an ethical audit (internal or external)
   - ☐ Test with diverse users to ensure inclusivity
   - ☐ Verify compliance with relevant standards (e.g., IEEE 7000-2021)
   - ☐ Iterate design based on trust-related feedback
6. Deployment Stage
   - ☐ Plan pilot or phased deployment approach
   - ☐ Prepare human-led orientation explaining the system's goals and operations
   - ☐ Create mentorship models (tech champions for peer training)
   - ☐ Design built-in tutorials or AI-guided onboarding
   - ☐ Start with low-stakes tasks before progressing to critical functions
   - ☐ Communicate clearly about data usage and privacy controls
7. Operation and Maintenance Stage
   - ☐ Implement continuous monitoring of system performance and user feedback
   - ☐ Create dashboards showing system health and anomalies
   - ☐ Program AI to defer to humans in novel situations
   - ☐ Schedule regular training refreshers and update communications
   - ☐ Implement predictive maintenance and AI model retraining
   - ☐ Monitor for signs of overreliance or skill atrophy
8. Feedback and Evolution Stage
   - ☐ Establish accessible feedback channels for users
   - ☐ Create process for reviewing feedback by development team or ethics committee
   - ☐ Schedule periodic reassessment of ethical risks, especially when scaling
   - ☐ Document any new scenarios or issues that emerge in real-world use
   - ☐ Update the system based on feedback and evolving ethical considerations
9. Measurement and Reporting
   - ☐ Track user trust levels through surveys or interviews
   - ☐ Monitor usage and override statistics
   - ☐ Log incidents and near misses
   - ☐ Track complaints or feedback tickets
   - ☐ Analyse diversity and fairness metrics
   - ☐ Measure adoption and retention rates
   - ☐ Evaluate performance against ethical constraints
   - ☐ Create transparent reporting dashboards
   - ☐ Establish response protocols for negative metrics

References

1. Yitmen, I.; Almusaed, A.; Alizadehsalehi, S. Investigating the causal relationships among enablers of the construction 5.0 paradigm: Integration of operator 5.0 and society 5.0 with human-centricity, sustainability, and resilience. *Sustainability* **2023**, *15*, 9105. [CrossRef]
2. Trstenjak, M.; Benešova, A.; Opetuk, T.; Cajner, H. Human Factors and Ergonomics in Industry 5.0—A Systematic Literature Review. *Appl. Sci.* **2025**, *15*, 2123. [CrossRef]
3. Bhatt, A.; Bae, J. Collaborative Intelligence to catalyze the digital transformation of healthcare. *NPJ Digit. Med.* **2023**, *6*, 177. [CrossRef]

4. Alves, M.; Seringa, J.; Silvestre, T.; Magalhães, T. Use of artificial intelligence tools in supporting decision-making in hospital management. *BMC Health Serv. Res.* **2024**, *24*, 1282. [CrossRef]
5. Ammeling, J.; Aubreville, M.; Fritz, A.; Kießig, A.; Krügel, S.; Uhl, M. An interdisciplinary perspective on AI-supported decision making in medicine. *Technol. Soc.* **2025**, *81*, 102791. [CrossRef]
6. Przegalinska, A.; Triantoro, T.; Kovbasiuk, A.; Ciechanowski, L.; Freeman, R.B.; Sowa, K. Collaborative AI in the workplace: Enhancing organizational performance through resource-based and task-technology fit perspectives. *Int. J. Inf. Manag.* **2025**, *81*, 102853. [CrossRef]
7. Loizaga, E.; Bastida, L.; Sillaurren, S.; Moya, A.; Toledo, N. Modelling and Measuring Trust in Human–Robot Collaboration. *Appl. Sci.* **2024**, *14*, 1919. [CrossRef]
8. Lykov, D.; Razumowsky, A. Industry 5.0 and Human Capital. In Proceedings of the International Scientific and Practical Conference "Environmental Risks and Safety in Mechanical Engineering", Rostov-on-Don, Russia, 1–3 March 2023; p. 2023.
9. Riar, M.; Weber, M.; Ebert, J.; Morschheuser, B. Can Gamification Foster Trust-Building in Human-Robot Collaboration? An Experiment in Virtual Reality. *Inf. Syst. Front.* **2025**, 1–26. [CrossRef]
10. Iqbal, M.; Lee, C.K.; Ren, J.Z. Industry 5.0: From Manufacturing Industry to Sustainable Society. In Proceedings of the 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Kuala Lumpur, Malaysia, 7–10 December 2022; pp. 1416–1421.
11. Shahruddin, S.; Sonet, U.N.; Azmi, A.; Zainordin, N. Traversing the complexity of digital construction and beyond through soft skills: Experiences of Malaysian architects. *Eng. Constr. Archit. Manag.* **2024**. [CrossRef]
12. Damaševičius, R.; Vasiljevas, M.; Narbutaitė, L.; Blažauskas, T. Exploring the impact of collaborative robots on human–machine cooperation in the era of Industry 5.0. In *Modern Technologies and Tools Supporting the Development of Industry 5.0*; CRC Press: Boca Raton, FL, USA, 2024.
13. Martini, B.; Bellisario, D.; Coletti, P. Human-Centered and Sustainable Artificial Intelligence in Industry 5.0: Challenges and Perspectives. *Sustainability* **2024**, *16*, 5448. [CrossRef]
14. Fraga-Lamas, P.; Fernández-Caramés, T.M.; Cruz, A.M.; Lopes, S.I. An Overview of Blockchain for Industry 5.0: Towards Human-Centric, Sustainable and Resilient Applications. *IEEE Access* **2024**, *12*, 116162–116201. [CrossRef]
15. Rame, R.; Purwanto, P.; Sudarno, S. Industry 5.0 and sustainability: An overview of emerging trends and challenges for a green future. *Innov. Green Dev.* **2024**, *3*, 100173. [CrossRef]
16. Brückner, A.; Wölke, M.; Hein-Pensel, F.; Schero, E.; Winkler, H.; Jabs, I. Assessing industry 5.0 readiness—Prototype of a holistic digital index to evaluate sustainability, resilience and human-centered factors. *Int. J. Inf. Manag. Data Insights* **2025**, *5*, 100329. [CrossRef]
17. Chew, Y.C.; Mohamed Zainal, S.R. A Sustainable Collaborative Talent Management Through Collaborative Intelligence Mindset Theory: A Systematic Review. *Sage Open* **2024**, *14*, 1–22. [CrossRef]
18. Rijwani, T.; Kumari, S.; Srinivas, R.; Abhishek, K.; Iyer, G.; Vara, H.; Gupta, M. Industry 5.0: A review of emerging trends and transformative technologies in the next industrial revolution. *Int. J. Interact. Des. Manuf.* **2025**, *19*, 667–679. [CrossRef]
19. Langås, E.F.; Zafar, M.H.; Sanfilippo, F. Exploring the synergy of human-robot teaming, digital twins, and machine learning in industry 5.0: A step towards sustainable manufacturing. *J. Intell. Manuf.* **2025**, 1–24. [CrossRef]
20. Abdel-Basset, M.; Mohamed, R.; Chang, V. A Multi-criteria decision-making Framework to evaluate the impact of industry 5.0 technologies: Case Study, lessons learned, challenges and future directions. *Inf. Syst. Front.* **2024**, 1–31. [CrossRef]
21. Hassan, M.; Zardari, S.; Farooq, M.; Alansari, M.; Nagro, S. Systematic Analysis of Risks in Industry 5.0 Architecture. *Appl. Sci.* **2024**, *14*, 1466. [CrossRef]
22. Karadayi-Usta, S. An Interpretive Structural Analysis for Industry 4.0 Adoption Challenges. *IEEE Trans. Eng. Manag.* **2020**, *67*, 973–978. [CrossRef]
23. Hsu, C.-H.; Li, Z.-H.; Zhuo, H.-J.; Zhang, T.-Y. Enabling Industry 5.0-Driven Circular Economy Transformation: A Strategic Roadmap. *Sustainability* **2024**, *16*, 9954. [CrossRef]
24. Carayannis, E.G.; Kafka, K.I.; Kostis, P.C.; Valvi, T. Robust, Resilient and Remunerative (R3) SMEs Ecosystems in the Quintuple Helix Context: Industry 5.0, Society 5.0 and AI Modalities Challenges and Opportunities for Theory, Policy and Practice. In *The Economic Impact of Small and Medium-Sized Enterprises*; Springer Nature: Berlin/Heidelberg, Germany, 2024; pp. 173–191.
25. Santos, B.; Costa, R.L.; Santos, L. Cybersecurity in Industry 5.0: Open Challenges and Future Directions. In Proceedings of the 21st Annual International Conference on Privacy, Security and Trust (PST), Sydney, Australia, 28–30 August 2024; pp. 1–6.
26. Chaudhuri, A.; Behera, R.K.; Bala, P.K. Factors impacting cybersecurity transformation: An Industry 5.0 perspective. *Comput. Secur.* **2025**, *150*, 104267. [CrossRef]
27. Ghobakhloo, M.; Mahdiraji, H.A.; Iranmanesh, M.; Jafari-Sadeghi, V. From Industry 4.0 digital manufacturing to Industry 5.0 digital society: A roadmap toward human-centric, sustainable, and resilient production. *Inf. Syst. Front.* **2024**, 1–33. [CrossRef]
28. Chrifi-Alaoui, C.; Bouhaddou, I.; Benabdellah, A.C.; Zekhnini, K. Industry 5.0 for Sustainable Supply Chains: A Fuzzy AHP Approach for Evaluating the adoption Barriers. *Procedia Comput. Sci.* **2025**, *253*, 2645–2654. [CrossRef]

29. Campagna, G.; Lagomarsino, M.; Lorenzini, M.; Chrysostomou, D.; Rehm, M.; Ajoudani, A. Promoting Trust in Industrial Human-Robot Collaboration Through Preference-Based Optimization. *IEEE Robot. Autom. Lett.* **2024**, *9*, 9255–9262. [CrossRef]
30. Chen, N.; Liu, X.; Hu, X. Effects of robots' character and information disclosure on human–robot trust and the mediating role of social presence. *Int. J. Soc. Robot.* **2024**, *16*, 811–825. [CrossRef]
31. Thurzo, A. How is AI Transforming Medical Research, Education and Practice? Bratisl. *Med. J.* **2025**, *126*, 243–248. [CrossRef]
32. Textor, C.; Zhang, R.; Lopez, J.; Schelble, B.G.; McNeese, N.J.; Freeman, G.; Visser, E.J. Exploring the Relationship Between Ethics and Trust in Human–Artificial Intelligence Teaming: A Mixed Methods Approach. *J. Cogn. Eng. Decis. Mak.* **2022**, *16*, 252–281. [CrossRef]
33. IBM. What Is AI Ethics? Available online: https://www.mckinsey.com/capabilities/quantumblack/our-insights/building-ai-trust-the-key-role-of-explainability#/ (accessed on 15 January 2025).
34. Giovine, C.; Roberts, R.; Pometti, M.; Bankhwal, M. Building AI Trust: The Key Role of Explainability. 2024. Available online: https://www.ibm.com/think/topics/ai-ethics (accessed on 20 February 2025).
35. Hadzovic, S.; Mrdovic, S.; Radonjic, M. A Path Towards an Internet of Things and Artificial Intelligence Regulatory Framework. *IEEE Commun. Mag.* **2023**, *61*, 90–96. [CrossRef]
36. Kolesnikov, M.; Lossi, L.; Alberti, E.; Atmojo, U.D.; Vyatkin, V. Addressing Privacy and Security Challenges at the Industry 5.0 Human-Intensive and Highly Automated Factory Floor. In Proceedings of the IECON 2024 50th Annual Conference of the IEEE Industrial Electronics Society, Chicago, IL, USA, 3–6 November 2024; pp. 1–6.
37. Malatji, M. Evaluating Human-Machine Interaction Paradigms for Effective Human-Artificial Intelligence Collaboration in Cybersecurity. In Proceedings of the 2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Bali, Indonesia, 17–19 December 2024; pp. 1268–1272.
38. Usmani, U.A.; Happonen, A.; Watada, J. Human-Centered Artificial Intelligence: Designing for User Empowerment and Ethical Considerations. In Proceedings of the 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Istanbul, Turkey, 8–10 June 2023; pp. 1–7.
39. Samarawickrama, M. The Irreducibility of Consciousness in Human Intelligence: Implications for AI, Legal Accountability, and the Human-in-the-Loop Approach. In Proceedings of the 2024 IEEE Conference on Engineering Informatics (ICEI), Melbourne, Australia, 20–28 November 2024; pp. 1–7.
40. D'souza, C.; Tapas, P. Diversity 5.0 framework: Managing innovation in Industry 5.0 through diversity and inclusion. *Eur. J. Innov. Manag.* **2024**. [CrossRef]
41. Aydin, E.; Rahman, M.; Bulut, C.; Biloslavo, R. Technological Advancements and Organizational Discrimination: The Dual Impact of Industry 5.0 on Migrant Workers. *Adm. Sci.* **2024**, *14*, 240. [CrossRef]
42. *ISO/TS 15066:2016*; Robots and Robotic Devices—Collaborative robots. ISO: Geneva, Switzerland, 2016.
43. *ISO10218-1:2025*; Robotics—Safety Requirements Part 1: Industrial Robots. ISO: Geneva, Switzerland, 2025.
44. *ISO10218-2:2025*; Robotics—Safety Requirements Part 2: Industrial Robot Applications and Robot Cells. ISO: Geneva, Switzerland, 2025.
45. Makeshkumar, M.; Sasi Kumar, M.; Anburaj, J.; Ramesh Babu, S.; Vembarasan, E.; Sanjiv, R. Role of Cobots and Industrial Robots in Industry 5.0. In *Intelligent Robots and Cobots: Industry 5.0 Applications*; Wiley Online Library: Hoboken, NJ, USA, 2025; pp. 43–63.
46. Liao, S.; Lin, L.; Chen, Q. Research on the acceptance of collaborative robots for the industry 5.0 era–The mediating effect of perceived competence and the moderating effect of robot use self-efficacy. *Int. J. Ind. Ergon.* **2023**, *95*, 103455. [CrossRef]
47. Golgeci, I.; Ritala, P.; Arslan, A.; McKenna, B.; Ali, I. Confronting and alleviating AI resistance in the workplace: An integrative review and a process framework. *Hum. Resour. Manag. Rev.* **2025**, *35*, 101075. [CrossRef]
48. Jacob, F.; Grosse, E.H.; Morana, S.; König, C.J. Picking with a robot colleague: A systematic literature review and evaluation of technology acceptance in human–robot collaborative warehouses. *Comput. Ind. Eng.* **2023**, *180*, 109262. [CrossRef]
49. Lawton, R.; Boswell, S.; Crockett, K. The GM AI Foundry: A Model for Upskilling SME's in Responsible AI. In Proceedings of the 2023 IEEE Symposium Series on Computational Intelligence (SSSCI), Mexico City, Mexico, 5–8 December 2023; pp. 1781–1787.
50. OECD. OECD Principles on Artificial Intelligence; OECD Legal Instruments: 2019. Available online: https://www.oecd.org/en/topics/sub-issues/ai-principles.html (accessed on 10 December 2024).
51. Directorate-General for Communications Networks, Content and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*; European Commission: Brussels, Belgium; Luxembourg, 2020.
52. Bohr, B. How to Turn Ethical Values Into System Requirements: Lessons Learned from Adopting a New IEEE Standard in the Business World. *IEEE Softw.* **2025**, *42*, 9–16. [CrossRef]
53. Brey, P.; Dainow, B. Ethics by design for artificial intelligence. *AI Ethics* **2024**, *4*, 1265–1277. [CrossRef]
54. Callari, T.C.; Segate, R.V.; Hubbard, E.M.; Daly, A.; Lohse, N. An ethical framework for human-robot collaboration for the future people-centric manufacturing: A collaborative endeavour with European subject-matter experts in ethics. *Technol. Soc.* **2024**, *78*, 102680. [CrossRef]

55. Palumbo, G.; Carneiro, D.; Alves, V. Objective metrics for ethical AI: A systematic literature review. *Int. J. Data Sci. Anal.* **2024**, 1–21. [CrossRef]
56. Thurzo, A. ProvableAI Ethics andExplainability in Medical andEducationalAIAgents: TrustworthyEthical Firewall. *Electronics* **2025**, *14*, 1294. [CrossRef]
57. Nurock, V.; Chatila, R.; Parizeau, M.H. What Does "Ethical by Design" Mean? In *Reflections on Artificial Intelligence for Humanity*; Braunschweig, M.I.B., Ghallab, M., Eds.; Springer Nature: Berlin/Heidelberg, Germany, 2021.
58. Drev, M.; Delak, B. Conceptual Model of Privacy by Design. *J. Comput. Inf. Syst.* **2021**, *62*, 888–895. [CrossRef]
59. Andrade, V.C.; Gomes, R.D.; Reinehr, S.; Freitas, C.O.; Malucelli, A. Privacy by design and software engineering: A systematic literature review. In Proceedings of the XXI Brazilian Symposium on Software Quality, Curitiba, Brazil, 7–10 November 2022; pp. 1–10.
60. Kassem, J.A.; Müller, T.; Esterhuyse, C.A.; Kebede, M.G.; Osseyran, A.; Grosso, P. The EPI framework: A data privacy by design framework to support healthcare use cases. *Future Gener. Comput. Syst.* **2025**, *165*, 107550. [CrossRef]
61. Chen, Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 567. [CrossRef]
62. Friedman, B.; Kahn, P.; Borning, A.; Huldtgren, A. Value Sensitive Design and Information Systems. In *Early Engagement and New Technologies: Opening up the Laboratory*; Doorn, N., Schuurbiers, D., Poel, I., Gorman, M., Eds.; Philosophy of Engineering and Technology; Springer: Berlin/Heidelberg, Germany, 2013.
63. Tolmeijer, S.; Christen, M.; Kandul, S.; Kneer, M.; Bernstein, A. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022; pp. 1–17.
64. Frigo, G.; Marthaler, F.; Albers, A.; Ott, S.; Hillerbrand, R. Training responsible engineers. Phronesis and the role of virtues in teaching engineering ethics. *Australas. J. Eng. Educ.* **2021**, *26*, 25–37. [CrossRef]
65. Yagoda, R.; Gillan, D. You Want Me to Trust a ROBOT? The Development of a Human–Robot Interaction Trust Scale. *Int. J. Soc. Robot.* **2012**, *4*, 235–248. [CrossRef]
66. Merritt, S.M.; Ilgen, D.R. Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Hum. Factors* **2008**, *50*, 194–210. [CrossRef] [PubMed]
67. Reijers, W.; Koidl, K.; Lewis, D.; Pandit, H.; Gordijn, B. Discussing Ethical Impacts in Research and Innovation: The Ethics Canvas. In *This Changes Everything—ICT and Climate Change: What Can We Do?* Kreps, D., Ess, C., Leenen, L., Kimppa, K., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 537, pp. 299–313.
68. Endmann, A.; Keßner, D. User Journey Mapping—A Method in User Experience Design. *I-Com* **2016**, *15*, 105–110. [CrossRef]
69. Berx, N.; Adriaensen, A.; Decré, W.; Pintelon, L. Assessing System-Wide Safety Readiness for Successful Human–Robot Collaboration Adoption. *Safety* **2022**, *8*, 48. [CrossRef]
70. Ong, J.C.; Chang, S.Y.; William, W.; Butte, A.J.; Shah, N.H.; Chew, L.S.; Ting, D.S. Ethical and regulatory challenges of large language models in medicine. Lancet Digit. *Health* **2024**, *6*, e428–e432. [CrossRef]
71. UNESCO. *Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2023.
72. Cancela-Outeda, C. The EU's AI act: A framework for collaborative governance. *Internet Things* **2024**, *27*, 101291. [CrossRef]
73. Rozenblit, L.; Price, A.; Solomonides, A.; Joseph, A.L.; Srivastava, G.; Labkoff, S.; Quintana, Y. Towards a Multi-Stakeholder process for developing responsible AI governance in consumer health. *Int. J. Med. Inf.* **2025**, *195*, 105713. [CrossRef]
74. Triandis, H.C.; Gelfand, M.J. Converging measurement of horizontal and vertical individualism and collectivism. *J. Pers. Soc. Psychol.* **1998**, *74*, 118–128. [CrossRef]
75. Pote, T.R.; Asbeck, N.V.; Asbeck, A.T. The ethics of mandatory exoskeleton use in commercial and industrial settings. *IEEE Trans. Technol. Soc.* **2023**, *4*, 302–313. [CrossRef]
76. Directorate-General for Research and Innovation. Industry 5.0, a transformative vision for Europe. *ESIR Policy Brief* **2021**. [CrossRef]
77. European Parliament, European Union Commission. 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* **2024**, 1–144. Available online: http://data.europa.eu/eli/reg/2024/1689/oj (accessed on 15 December 2024).
78. NIST. *Artificial Intelligence Risk Management*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
79. Office of Science and Technology Policy. Blueprint for an AI Bill of Rights. The White House. 2022. Available online: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/ (accessed on 15 November 2024).
80. *IEEE 7001-2021*; IEEE Standard for Transparency of Autonomous Systems. IEEE: New York City, NY, USA, 2024.
81. *IEEE 7007-2021*; IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems. IEEE: New York City, NY, USA, 2021.
82. *ISO/IEC 42001:2023*; ISO/IEC Information Technology—Artificial Intelligence—Management System. ISO: Geneva, Switzerland, 2023.

83. Crootof, R.; Kaminski, M.; Price, W.N. Humans in the Loop. *Vanderbilt Law Rev.* **2023**, *76*, 429. [CrossRef]
84. Ramírez-Moreno, M.; Carrillo-Tijerina, P.; Candela-Leal, M.; Alanis-Espinosa, M.; Tudón-Martínez, J.; Roman-Flores, A.; Lozoya-Santos, J. Evaluation of a Fast Test Based on Biometric Signals to Assess Mental Fatigue at the Workplace—A Pilot Study. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11891. [CrossRef]
85. Moshawrab, M.; Adda, M.; Bouzouane, A.; Ibrahim, H.; Raad, A. Smart Wearables for the Detection of Occupational Physical Fatigue: A Literature Review. *Sensors* **2022**, *22*, 7472. [CrossRef]