information

Review

Collaborative Intelligence for Safety-Critical Industries: A Literature Review

Inês F. Ramos [1,*], Gabriele Gianini [2,*], Maria Chiara Leva [3] and Ernesto Damiani [1]

[1] Department of Computer Science, Università degli Studi di Milano, 20122 Milano, Italy; ernesto.damiani@unimi.it
[2] Department of Informatics, Systems and Communication (DISCo), Università degli Studi di Milano-Bicocca, 20126 Milano, Italy
[3] School of Food Science and Environmental Health, Technological University Dublin, D07 H6K8 Dublin, Ireland; mariachiara.leva@tudublin.ie
* Correspondence: ines.fernandes@unimi.it (I.F.R.); gabriele.gianini@unimib.it (G.G.)

Abstract: While AI-driven automation can increase the performance and safety of systems, humans should not be replaced in safety-critical systems but should be integrated to collaborate and mitigate each other's limitations. The current trend in Industry 5.0 is towards human-centric collaborative paradigms, with an emphasis on *collaborative intelligence* (CI) or Hybrid Intelligent Systems. In this survey, we search and review recent work that employs AI methods for collaborative intelligence applications, specifically those that focus on safety and safety-critical industries. We aim to contribute to the research landscape and industry by compiling and analyzing a range of scenarios where AI can be used to achieve more efficient human–machine interactions, improved collaboration, coordination, and safety. We define a domain-focused taxonomy to categorize the diverse CI solutions, based on the type of collaborative interaction between intelligent systems and humans, the AI paradigm used and the domain of the AI problem, while highlighting safety issues. We investigate 91 articles on CI research published between 2014 and 2023, providing insights into the trends, gaps, and techniques used, to guide recommendations for future research opportunities in the fast developing collaborative intelligence field.

Keywords: collaborative intelligence; AI; safety-critical industries

check for updates

Citation: Ramos, I.F.; Gianini, G.; Leva, M.C.; Damiani, E. Collaborative Intelligence for Safety-Critical Industries: A Literature Review. *Information* **2024**, *15*, 728. https://doi.org/10.3390/info15110728

Academic Editor: Dharmaraj Veeramani

Received: 3 October 2024
Revised: 2 November 2024
Accepted: 5 November 2024
Published: 12 November 2024

1. Introduction

Industry 4.0 was characterized by the increase in automation of processes and the adoption of artificial intelligence (AI) in a variety of industries, such as manufacturing, the process industry, healthcare, and the transport industry. AI-driven automation has certainly demonstrated its value by increasing productivity, reducing cost, and adding a new layer of safety; however, the rapid and widespread adoption of these intelligent systems without proper consideration of the ethical, societal, and safety implications has quickly brought to light some drawbacks.

Smart automated systems contribute to the overall system performance and safety by automating routine tasks, identifying and anticipating hazards and failure events, and supporting operator awareness and decision-making in high-complexity troubleshooting situations; even so, the human still has the important task of monitoring possible automation failures, carrying out manual procedures if needed, and performing decision-making. Failure in the proper integration between them has previously led to disastrous consequences (such as the accident of Air France flight 447 in 2009) due to operator loss of expertise, reduced vigilance and situational awareness, complacency, reduced adaptability, or information overload [1]. Particularly for safety-critical systems, such as those in transport, nuclear, process, and infrastructure industries, it is expressly dangerous to over-rely and overestimate the capabilities of intelligent systems. Human experts should

not be replaced but integrated into the loop so that humans and systems can collaborate dynamically to mitigate each other's limitations and enhance each other's strengths. Therefore, the current trend in Industry 5.0 is collaboration and cooperation between humans and intelligent systems to achieve optimal performance. This approach is not novel but, as Sendhoff and Wersing (2020) have noted in their work 'Cooperative Intelligence—A Humane Perspective' [2], a 'human-centric' view of AI-based systems that has recently been taken to the forefront.

Despite the potential benefits of human–machine collaboration for high-stakes industries, there are specific challenges to its adoption, such as the complexity of integrating human and machine/AI roles and feedback [3], safety issues related to low model transparency and explainability, vulnerability to adversarial attacks and human bias [4], and system safety and robustness certification [5]. Following the publication of the Ethics Guidelines for Trustworthy AI presented in 2020 by the High-Level Expert Group on Artificial Intelligence [6], the new essential health and safety requirements for machinery with AI (Machinery Regulation (EU) 2023/1230) and the EU AI Act (Regulation (EU) 2024/1689), major concerns were raised regarding the risks of harm to the health, safety, or fundamental rights of people in their interaction with autonomous systems/machines, particularly those with self-evolving behavior, with close interaction with humans, or those used as a safety component or product. Innovation and the use of state-of-the-art CI technology may be hampered by these regulatory constraints, low digitization levels, and lack of AI expertise, specifically for small and medium enterprises (SMEs) in industry [7].

Collaborative intelligence (CI) is a complex and interdisciplinary field, leveraging advances in AI and machine learning (ML) to achieve higher levels of synergy between humans and machines. As such, we aim to outline the large diversity of AI methods and collaboration paradigms explored in the literature. We intend to assist engineers and practitioners in the related fields of robotics, human–computer interaction (HCI), ergonomics, AI, safety engineering, and behavioral sciences, to narrow down CI solutions for domain-specific problems and reduce the barrier to adoption, by offering a broad overview of potential applications of collaborative intelligence, with safety in mind.

The systematic categorization proposed next is based on the main drive for the collaboration, either assistance to the human or to the machine, enabling the capture of a diversity of solutions and the discovery of how different AI methods can be employed for the same problem. Collaborative intelligence, also known as cooperative intelligence [2], or Hybrid Intelligent Systems [8], can be defined in different ways depending on the goals and modes of collaboration/interaction. Akata et al. (2020) [9] has defined hybrid intelligence as a combination of human and machine intelligence that enhances and leverages human capabilities, and achieves goals unattainable by the human or machine alone, through human–machine teaming. While in [2], a wider definition was presented, considering cooperative intelligence as a state of the system required to establish a relationship between multiple agents (human or not) and that beyond working together to achieve a common goal, it could be leveraged just to benefit from each other and live together synergistically. Here, we use a general interaction-based definition, where the human and the machine/algorithm intelligence are interdependent and interact to achieve a goal (contrary to being used independently for the same goal), identifying two main types of purposeful interactions in human–machine collaboration: the machine assists the human, or the human assists the machine.

In the human–machine collaboration scenario, in which the main goal is for the *machine to assist the human-in-the-loop* (HIL), the objective of the intelligent system is to monitor and support the human operator in their tasks:

- By monitoring the human and the task context and using the knowledge to detect critical conditions, support, or adapt to the operator's cognitive status for optimal system performance;

- By amplifying the operator's physical capabilities, as in the case of exoskeletons that can be worn by the user and enhance their physical performance, or telerobots that allow the human to perform a variety of complex tasks using the capabilities of robots;
- By embodying human physical capabilities as in the case of cobots, that can work alongside humans performing complementary tasks or substituting the operator in a dynamic way.

In the type of human–machine collaboration in which the main goal is for the *human-in-the-loop to assist the machine* during a learning process, a test process, or a deployment process, the human can provide assistance:

- By annotating data to be employed by AI algorithms. This step is crucial for the performance of supervised machine learning (ML); however, due to the effort and cost of manual labeling of large amounts of data, more efficient learning strategies have been developed, such as those using active learning that tries to maximize the model's performance while querying a human to annotate a data sample as few times as possible.
- By the direct demonstration of tasks, as in the robot learning from a demonstration paradigm or by direct intervention on the automated process.
- By using expert knowledge to validate and explain intelligent machine behavior that, despite the emergence of explainable AI techniques, is still required to ensure the outcomes and that the generated explanations match the expected behavior in a reliable and unbiased way.

2. Related Works

Safety in industry is regulated by legal requirements and standards, in particular for the development and deployment of systems that interact with humans. For instance, in human–robotic collaboration applications, the new essential health and safety requirements under the EU Machinery Regulation 2023/1230 apply and have to be considered during the design stage and throughout the lifecycle of the robotic system. A commonly applied standard is the harmonized standard EN ISO 12100:2010 [10], regarding risk assessment and risk reduction principles, described by three steps: (1) the implementation of inherent safe design measures that aim to eliminate or reduce the risk of hazards, (2) the implementation of safeguarding and/or complementary protective measures when a hazard cannot be eliminated or its risk is reduced at the design stage, and (3) if risks remain, the disclosure of information for use that shall recommend safety procedures to be implemented by the user. A more specific harmonized standard for the context human–robot collaboration is the EN ISO 10218-2:2011 [11], including safety requirements and protective measures.

For the safety of railway systems, specifically for control software with impact on safety, the recent EN 50716:2023 [12] provides lifecycle development processes and technical requirements, including formal methods and the integration of ML/AI techniques.

In the case of highly automated vehicles, the complexity of the environment and intended functionality requires safety guidance on multiple system levels. General functional safety can be addressed by the requirements of EN ISO 26262:2018 [13], while SOTIF EN ISO 21448:2022 [14] focuses on design, verification and validation guidance for the specification of the intended functionality, which can apply to hazards caused by the limitations of the implemented ML/AI in advanced driver-assistance systems. In the case of fully autonomous vehicles, the standard UL 4600 (2023) [15] proposes an approach for assessing and validating system safety, while for systems requiring human interaction and supervision, other standards specific to safe and ergonomic design should be applied (such as EN ISO 9241-210:2019 [16]).

However, with the fast development and introduction to the market of CI technology, other current legislation and standards may require updating to deal with the emergent challenges of new digital technologies, such as the artificial intelligence, Internet of Things and robotics domains. To support the operationalization of compliance to new safety

requirements, in addition to leveraging existing and emerging standards, state-of-the-art research efforts can shed light on domain-specific solutions.

Previous surveys and literature reviews have approached some of these topics in a more focused manner, mostly covering two main general lines of research: intelligent human–machine interactions, namely human–robot and human–computer interactions, and human-in-the-loop learning.

Hua et al. (2021) [17] surveyed state-of-the-art deep learning, reinforcement, imitation, and transfer learning AI methods for robot control and adaptation to diverse complex environments and tasks, including their application to human–robot collaboration. In the recent work of Borboni et al. (2023) [18], the potential role of AI in the use of cobots for industrial applications was explored, and the state-of-the-art research on AI-based collaborative robotic applications was analyzed. Liu and Wang (2018) [19] also covered human–robot collaboration in their review, specifically, gesture recognition used for communication between human workers and robots. The work reports on the most important technologies and algorithms of gesture recognition existing in the current research. A model and classification scheme of gesture recognition for human–robot collaboration is proposed, with four technical components: sensor technologies, gesture identification, gesture tracking, and gesture classification.

Another possible effective communication channel between robots, machines, and humans is the human gaze. Zhang et al. (2020) [20] performed a literature review of human gaze modeling and its potential applications. Human gaze data can be used by AI to develop the intelligent attentional selection of information, or for the AI agents to be aware of the human cognitive and emotional state, fostering more natural communication and interactions between them. The work reviewed human-gaze-assisted AI agents in multiple fields, such as computer vision, natural language processing, imitation, and reinforcement learning, as well as robotics. In a more high-level and broad mapping study by Šumak et al. (2022) [21], state-of-the-art AI methods for intelligent human–computer interaction using sensor data were reviewed. The mapping found that studies have mostly focused on recognizing the emotion and stress of HCI users, followed by gestures and facial expressions identification, using, more frequently, deep learning algorithms, including CNNs, and from the classical machine learning algorithms, SVMs.

An alternative type of collaboration is human-in-the-loop learning, typically employing the human to deal with sparse data, lack of training data, or improve the performance of machine learning methods with expert task knowledge. Wu et al. (2022) [22] analyzed the literature from a data perspective, categorizing methods based on the stage at which the human was added to the loop—in the data processing stage, model training stage, or in the design and application of the system. From the point of view of who is in control during learning, human-in-the-loop machine learning methods can be mainly divided into active learning, where the system is responsible for the learning process and interacts with the human for data annotation; interactive machine learning, where the interaction is less structured, more frequent, and incremental; and machine teaching, where the human expert is the responsible for the learning process by transferring knowledge [23]. From a safety perspective, human–machine collaboration has been leveraged for safe learning and anomaly detection by incorporating human knowledge, demonstration, supervision, or feedback in the learning process [24]. Due to the black-box nature of AI/DL-based systems and their weakness to adversarial attacks and out-of-distribution inputs [25], the unique skills of humans such as pattern discrimination, high-level conceptualization, and hazard identification can be used by the system to mitigate safety risks.

To our knowledge, the current literature explorations of collaborative intelligence and AI problems, techniques, and challenges have not taken into account the common aspects between the previously mentioned disparate topics of work and how collaborative interactions are at the core of the targeted technologies. With this survey, we intend to give a general and unified overview of recent works employing AI methods for collaborative intelligence problems, with an additional focus on safety or application to safety-critical

industries. For these industries, safety is put on the forefront of performance; however, both the probabilistic nature of AI and the uncertainty of human behavior make the regulatory certification of CI methods a challenge [24]. The work of [5] provided a general overview of potential safety issues in human–machine collaboration from the machine side, the human side, and the interaction side, and possible countermeasures to be considered early on in the development of human–machine teaming, lacking, however, a direct link to specific tasks and application domains. Safety is an emergent property of the combination of components of a system [24], and as such, the risk assessment process should involve the setting of the operating conditions and identification of potential hazards for the individual components according to their role in the system, highly dependent on the application domain. Our aim is to promote the progress and implementation of collaborative intelligence in safety-critical industries by providing a wide array of application examples, their limitations, and insight into outstanding safety concerns. We analyze the trends and gaps in the AI problems addressed, the interaction tasks, and techniques used in the latest collaborative intelligence research landscape to set the directions for future research on safer human–machine collaboration. Furthermore, we uncover from the reviewed papers a sub-categorization of collaborative intelligence tasks (Figure 1) and link it to the previously identified types of CI interactions (Machine assists human-in-the-loop or human-in-the-loop assists the machine).

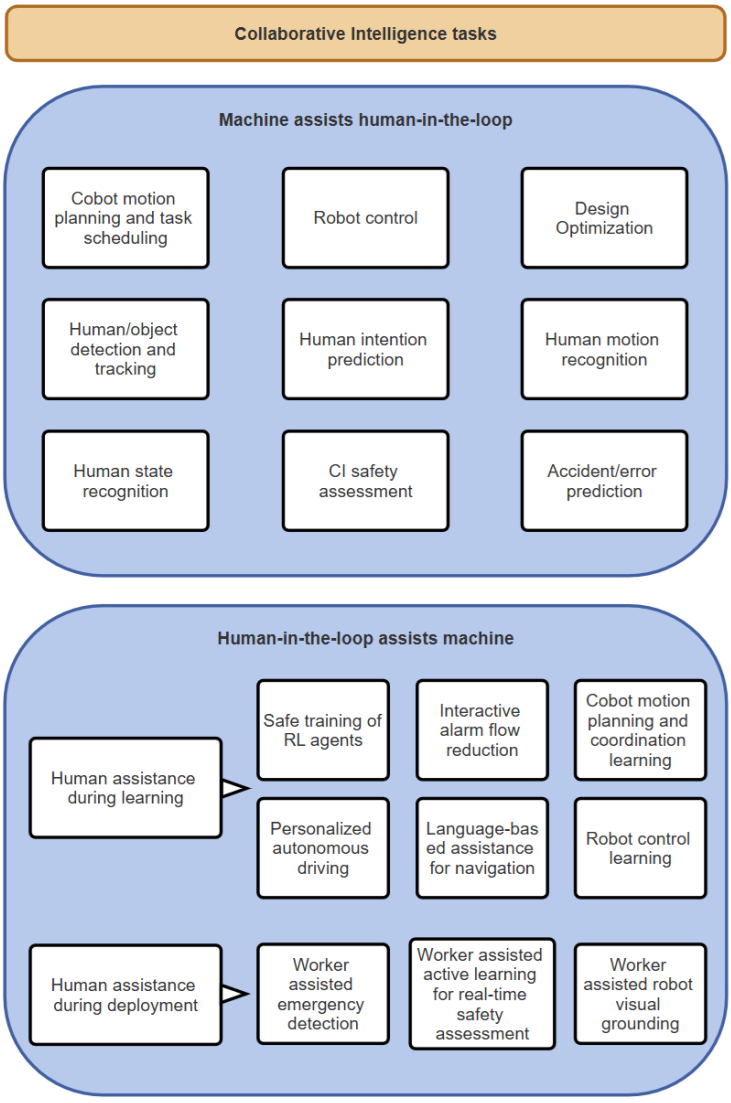

Figure 1. CI interaction types and task categories.

As a wide survey of AI methods for CI solutions, we cover a broad range of techniques that aim to emulate human intelligence through computer algorithms [26]. AI techniques generally fit into three main groups: machine reasoning, machine learning, and robotics [27]. These three groups are, however, a very coarse classification, blending categorizations of learning style, similarity in form or function, and the AI problem it tackles. For this review, a more comprehensive taxonomy will be used expanding on the AI Knowledge Map developed by Corea (2019) [28]. We categorize AI solutions based on the intelligence problem that is solved (AI problem domain) and the type of AI approach used (AI paradigms).

Below, we summarize the standard domain categories that *AI problems* typically fall into:

- Perception domain : Tasks of sensing and understanding signals from the physical world, transforming those data into valuable and relevant information for the specific application.
- Reasoning domain: Tasks that use logic and existing knowledge to solve new problems.
- Knowledge representation domain: Tasks that try to represent knowledge in a way that can be manipulated and understood.
- Planning domain: Tasks that try to find the best way to achieve goals within certain spatial, temporal, or resource constraints.
- Communication domain: Tasks regarding the understanding of language and communication.
- Control domain: Tasks regarding the monitoring and regulation of a system to achieve a desired result.

Next, we specify the categories of *AI paradigms* that will be employed:

- Logic-based approaches: Approaches that use logic and abstract high-level representations of knowledge to solve problems. Logic-based approaches are the foundation of classical symbolic AI and are used for three main problems: knowledge representation, reasoning, and model-checking and verification [29].
- Knowledge-based approaches: Approaches that use knowledge representation systems, such as large knowledge bases, to represent explicitly and declaratively known notions, information, and rules in order to infer new implicit symbolic knowledge [30].
- Probabilistic approaches: Approaches that use probabilistic knowledge and scenarios, such as Naïve Bayesian Networks, Markov Models, and Restricted Boltzmann Machines [31].
- Machine learning approaches: Approaches that allow a machine/computer to learn from data, i.e., data-driven methods, such as Artificial Neural Networks and Ensemble Learning algorithms [32,33]. These approaches can be sub-classified based in the degree of supervision they have during learning (unsupervised, supervised, or semi-supervised learning) or the prior knowledge (inductive, transductive or transfer learning).
- Neuro-symbolic approaches: Approaches that integrate symbolic and sub-symbolic models at any stage of an intelligent system (model design, input, output, reasoning, or learning stage) [34].
- Search and Optimization approaches: Approaches that allow to intelligently search through many solutions to achieve the target objective function value while satisfying the problem's constraints [35]. In general, optimization methods can be divided into either local or global algorithms.
- Embodied intelligence approaches: Approaches that allow for an agent to have higher intelligence, such as movement, perception, interaction, and visualization abilities. Approaches like reinforcement learning and learning (programming) from demonstration include not only the intelligent agent but also its "body" that interacts with the world according to some constraints, and the specific environment it is situated in. A broader definition can be taken to include simulation, where the embodied intelligence

agent can purposefully exchange energy and information with a simulated physical environment [36].

Section 3 contains the reviewed works organized according to the general type of CI interaction, the sub-categorization of CI tasks found during the analysis (displayed in Figure 1), and the AI taxonomy mentioned above. In addition, a summary of the papers is given for each CI task sub-category. Lastly, a discussion of the most important insights discovered through the review and analysis of the research questions is presented in Section 4, also providing recommendations and future research opportunities for the field of collaborative intelligence.

3. Results

To conduct the review, the following databases were searched and used as the principal research systems: ACM digital library, ISI Web of Knowledge, Wiley Inter Science, Scopus, and IEEE Xplore. Works published between 2006 and 2023, in conferences of class A or journals of the first or second quartile, related to the main topics of the review were identified: collaborative intelligence, artificial intelligence, and safety. Articles not related to industrial applications or safety-critical industries were not included.

After the definition of the classification strategy, the papers were categorized according to the general CI task sub-category, the AI paradigm employed, the AI problem domain addressed, the collaborative intelligence goal achieved, and the application domain (if mentioned). In addition, the research goals, research methodology, and results were analyzed for each category. To guide the discovery of trends and gaps in the field, we studied the type of CI tasks addressed and the common AI methods and techniques employed for collaborative intelligence problems, the industries most targeted by the research, how the safety of human–machine interactions is addressed, and identified common future lines of research/work.

The final 91 articles retrieved from the literature are presented and described by CI interaction type and in chronological order in Tables 1–11. From an initial analysis, it was observed that most of the articles retrieved that fit our focused research topics were published recently, between 2019 and 2023 (Figure 2), while no article published between 2006 and 2014 was found that matched our search queries. The most addressed collaborative interactions involved the machine assisting the human with human–robot collaboration tasks in the manufacturing industry domain. The second largest portion of the papers provided AI solutions for CI tasks independently of the application domain, not having specified it in the paper, or mentioning only potential target applications. However, the survey was able to retrieve research works focused on safety-critical industries, such as automotive, maritime, aviation, nuclear, and rail industries (Figure 3).

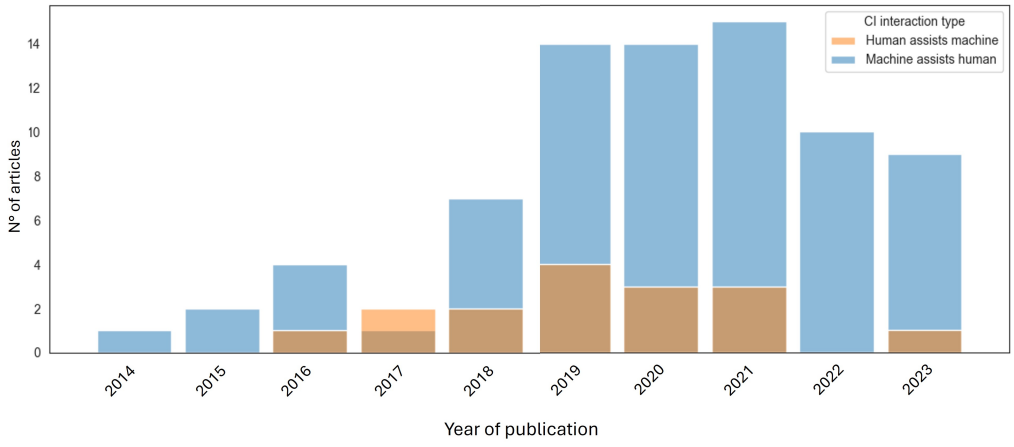

Figure 2. Year of publication of articles using AI for CI grouped by CI interaction type.

The distribution of the papers following the defined AI problem and the paradigm taxonomy is depicted in Figures 4 and 5, respectfully. A tendency is clear for the type of AI problems that are approached in collaborative intelligence research, with most of the articles including perception, planning, and/or control problems. In terms of AI paradigms, machine learning is the most popular technique in the retrieved papers, followed by probabilistic approaches. It is worthy of note that a large portion of work has employed either hybrid techniques or made use of more than one type of AI method to solve multiple AI problems, often necessary to achieve complex CI solutions.

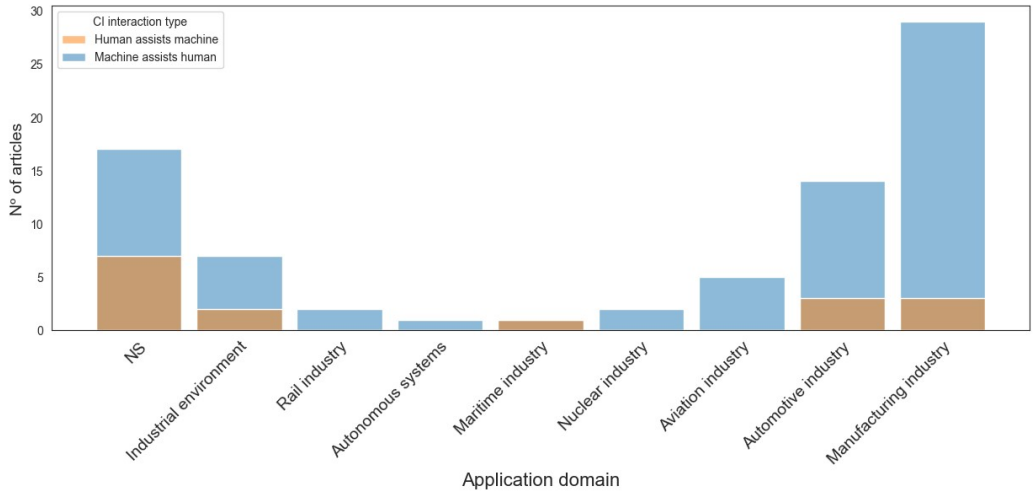

Figure 3. Domain of application of the papers.

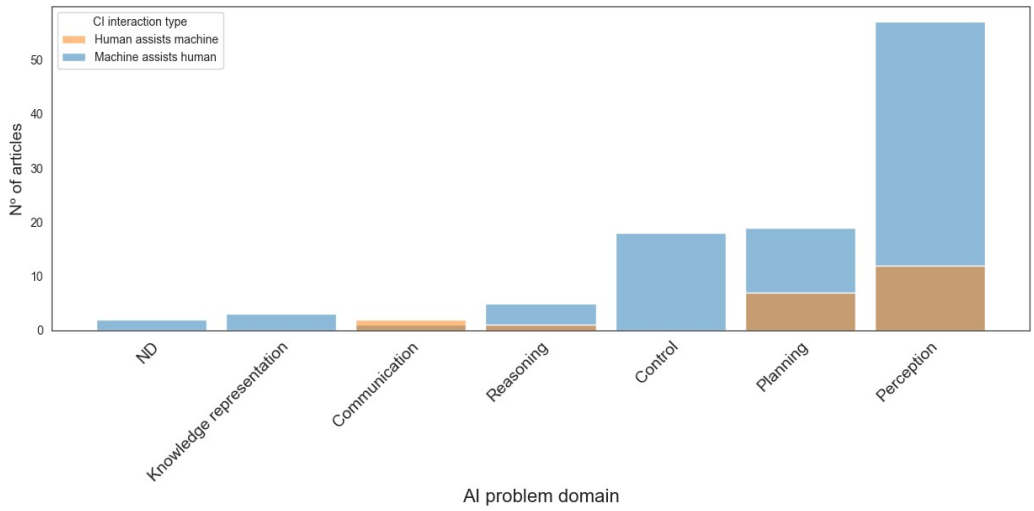

Figure 4. Number of articles for each AI problem domain category.

In the selected research works, a variety of common collaborative intelligence tasks were found (Figure 6). The identified subtypes of CI tasks in which the machine assists the human were *Cobot motion planning and task scheduling*, *Robot control*, *Design optimization*, *Human/object detection and tracking*, *Human intention prediction and motion recognition*, *Human state recognition*, *Human mental model estimation*, *CI safety assessment*, and *Accident/error prediction*; the tasks related to humans assisting the machine were *Human assistance during learning*, including the safe training of RL agents, interactive alarm flood reduction, personalized autonomous driving, language-based assistance for navigation, cobot motion planning and coordination learning, and robot control learning, and *Human assistance during deployment*, addressing worker-assisted emergency detection, worker-assisted active learning for real-time safety assessment, and worker-assisted robot visual grounding.

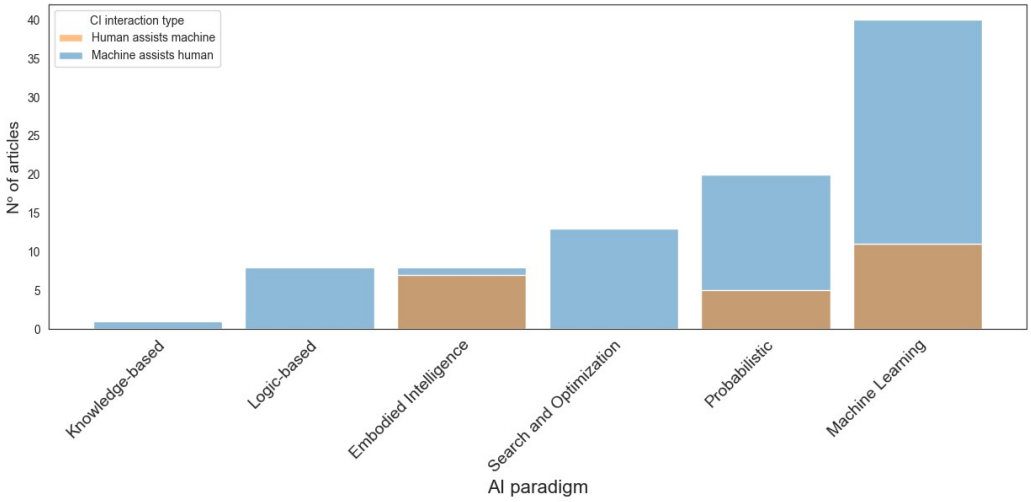

Figure 5. Number of articles for each AI paradigm category.

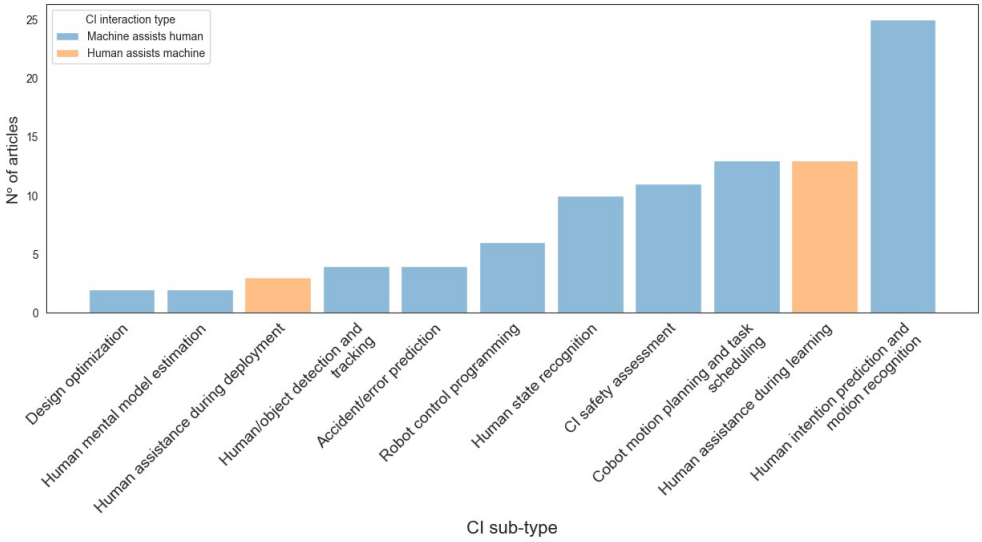

Figure 6. Number of articles for each CI task sub-category.

3.1. Machine Assists Human-in-the-Loop

3.1.1. Cobot Motion Planning and Task Scheduling

Robot and cobot (collaborative robots designed to share the same workspace with human workers) motion planning and task scheduling require the system to be aware of the human, to avoid collision, and coordinate tasks between them, without loss of speed or efficiency. Different AI methods have been used in the literature to implement robot motion planning tasks or human–robot task scheduling, which aim at assisting the human and achieving optimal coordination and system performance in collaborative tasks. Within the aim of this survey, thirteen works have been retrieved that employ AI techniques for this type of CI task and interaction (Table 1).

Table 1. Description of the retrieved articles that use AI for CI tasks related to cobot motion planning and task scheduling, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Cobot motion planning and task scheduling	[37]	Robot path and task planning and scheduling	Planning	Probabilistic method for HOV and execution time computation, RRT for trajectory planning and flexible timeline-based task planning and scheduling	Probabilistic, Search and Optimization; Logic-based	Manufacturing industry
	[38]	Robot task planning and scheduling	Control, Planning	Iterative IRL for human preference learning	Search and Optimization	Autonomous systems
	[39]	Robot motion and task planning	Planning, Control	Flexible timeline-based task planning model	Logic-based	Manufacturing industry
	[40]	Robot motion/path planning	Perception, Planning	Kalman filter for velocity estimation and RRT with obstacle velocity-based costmap for path generation	Probabilistic, Search and Optimization	Manufacturing industry
	[41]	Learn human expected robot behavior for generation of explicable robot plans	Planning	Regression models to learn the mapping of distance between agent's and humans' expected plan to explicability score and anytime greedy search to progressively search for plans with better explicability scores	Machine Learning, Search and Optimization	NS
	[42]	Robot motion/path planning	Perception, Planning	Probabilistic movement primitives, GMM	Probabilistic	NS
	[43]	Robot motion/path planning	Planning, Control	Deep RL	Embodied Intelligence	Manufacturing industry
	[44]	Learning of desired driver's path for semi-autonomous driving	Control task	Iterative learning control	Search and Optimization	Automotive industry
	[45]	Intention-guided semi-autonomous driving for mobile robot teleoperation	Planning	BIAS for path generation	Search and Optimization	NS
	[46]	Robot motion/path planning	Planning	Feedforward ANN	Machine learning, Embodied Intelligence	Manufacturing industry
	[47]	Robot task planning and scheduling	Planning, Control	Double DDPG	Embodied Intelligence	Manufacturing industry
	[48]	Robot motion and task planning	Perception, Planning	T-RRT and MPC methods for optimal tool delivery position, and trajectory and GMR for human motion prediction	Search and Optimization	Manufacturing industry
	[49]	Robot path planning	Perception, Planning	Artificial Potential Fields for collision avoidance	ND	NS

The acronyms used have the following expansions: Artificial Neural Network (ANN), Biased Incremental Action Sampling (BIAS), Gaussian Mixture Model (GMM), Gaussian Mixture Regression (GMR), Human Occupancy Volume (HOV), Inverse Reinforcement Learning (IRL), Model Predictive Control (MPC), Not Defined (ND), Not Specified (NS), Reinforcement Learning (RL), Rapidly Exploring Random Tree (RRT), Transition-based Rapidly Exploring Random Tree (T-RRT).

The main type of tasks identified within this sub-category were cobot motion path planning, cobot task planning and scheduling, and methods aimed at autonomous or semi-autonomous driving path learning that can also be translated into the cobot domain. These works contribute with innovative technology that aims to achieve one of the goals of Industry 5.0: seamless and safe human–machine interaction and collaboration to capitalize on the human worker's added value and the efficiency of advanced robotic solutions. For safe collaboration, earlier work defined Human Occupancy Volumes (HOVs) to constrain the path search space and compute the collision probability between human and robot [37], while more recent systems aimed at dynamic real-time collision avoidance using vision systems or models for human motion prediction [40,42,46,48]. For autonomous/semi-autonomous driving, of either teleoperated mobile robots [38,45] or vehicles [44], the goal of collaboration is instead the incorporation of the driver's intention/preference in the generated paths, while assisting the driver with automated path generation, compliance with safety constraints, and the monitoring of environmental obstacles.

Dynamic and flexible task planning and scheduling methods can leverage the available resources (robots and operators) in the optimal way, by dynamically allocating tasks in a specific sequence, constrained by their skills, temporal and synchronization limitations, and execution time uncertainty [37,39]. More recent works have addressed some of the

gaps in previous methods, namely, the inability to change the assembly plans according to changes in the environment [47] that may be required in manufacturing SMEs with shared resources between assemblies, and the weak awareness of humans' mental model and their expected robot behavior, a key ability for trust and reduction in safety risks [41]. Factors such as robot visibility and the comfort of the human's interaction position can also reduce the distance between the robot's and the humans' expected plan of the same task [48].

Different types of AI paradigms and methods have been applied to solve the complex tasks of robot motion planning and task scheduling, falling mostly under Search and Optimization paradigms for Control and Planning AI problems. We highlight the classical Rapidly Exploring Random Tree (RRT) method for cobot motion planning [37,40,48]. The RRT algorithm is a path search algorithm that performs the efficient sampling of high-dimensional spaces with algebraic constraints (due to obstacles) and differential constraints (constraints on the system configuration variables that derive from inaccessible paths for the system) by incrementally biasing the search tree towards the unexplored state space. It can be applied using different kinds of sensor data to provide information about human occupancy and collision risk, and it can be combined with a task scheduling framework to deal with temporal uncertainty caused by changes in the robot trajectory. More modern strategies can employ neural networks for situations that require speed and reliability, and switch to a more accurate and safe baseline deterministic algorithm, using a simplex architecture for the safety assurance of the path planning system [46]. In the case of highly dynamical and cluttered environments, with strict real-time control requirements for collision avoidance, ref. [49] proposed the traditional path-planning algorithm Artificial Potential Fields connected to a digital twin of the current environment. The algorithm does not require optimization and allows real-time reaction to environment changes by defining attractive forces towards a goal and repulsive forces from obstacles. The digital twin provides both location, orientation and distance information but can also communicate task-specific interaction parameters. Future work still requires nonetheless a method to update the digital twin for the detection of dynamic obstacles. For vehicle path generation instead, in considering the driver's preferences, iterative learning strategies were proposed as an alternative for the data-hungry Deep Reinforcement Learning strategies, to enable learning through repeated cooperation with the human driver [44,45].

Timeline-based task planners were the most commonly used methods for task planning and scheduling problems [37,39], modeling the human (state variable with uncontrollable values) and the robot behavior (state variable with partially controllable values) as multi-valued state variables to be controlled over time, based on causal, temporal, and synchronization legal constraints that dictate transitions between values and the duration of the value. The evolution of a multi-valued state variable over time is called the timeline of that state variable. A digital twin-based framework can also be developed to account for the target industrial environment, the optimization of assembly processes, and adaptation to environmental changes, employing reinforcement learning methods, such as the Double Deep Deterministic Policy Gradients (D-DDPGs) [47]. When the state space, or action space, has high dimensionality and is very complex (such as in the case of a digital twin of an industrial cell), deep neural networks can be used to either represent the system states, or to approximate the optimal Q function, the optimal policy function, or the model (state transition function) in an efficient way. The network can be trained by minimizing the difference between the expected reward and the real reward received by the environment. The combination of reinforcement learning and deep learning is called Deep RL, and has shown recently a lot of progress, even if with limited applicability to real-world problems so far due to the amount of data required to train the models and safety concerns, particularly for model-free RL, where the agent learns by trial and error.

Despite the fast development of motion planning and tasks scheduling technology, previous research has highlighted the important role of human autonomy in supporting the well-being of workers. A balance must be found between worker autonomy in task allocation and process control, and the use of these technologies to reduce the mental

demand on the workers [50]. Assistance systems should be supportive but not overly complex, avoiding fully autonomous decision-making [51]. From another perspective, intelligent systems can detect human decision-making bias, and select the most effective way to explain and persuade the worker to take the most beneficial action [5].

In cobot applications, where the operator and the robot share the same workspace, worker physical safety is also paramount. The safety measures taken by the proposed solutions commonly fall under two categories: dynamic path replanning or decrease in robot speed to avoid collisions. Both usually rely on probabilistic approaches to find the optimal balance between safety and efficiency. Human motion intention prediction can confer to the robot the adaptation and proactive planning ability to better assist the human operator; however, mutual understanding is required for effective interactions (value alignment) [5]. An explainability metric can be computed or trained to assess the generated cobot motion plans, promoting the situational awareness and trust of the worker.

In driver-assistance systems, safety is addressed by performing collision monitoring and generating safe paths. With RL-based solutions, there is the risk during the learning phase that the robot performs dangerous actions while interacting with the human and the environment. As such, the use of a digital twin to conduct the learning is recommended [47], supplemented with a method for robot action safety/risk evaluation.

3.1.2. Robot Control Programming

Robot control is an essential problem when humans collaborate with cobots or work with other types of robots, such as in teleoperation. Particularly in manufacturing tasks, cobots are required to perform dynamic complex tasks while maintaining performance, and ensuring the safety of the human collaborators. AI can be used for autonomous control, to manage the movement and behavior of robotic manipulators, providing safety solutions for automatic collision avoidance, path adaptation, and speed adaptation control, according to detected obstacles or predicted human motion. The AI is subsequently assisting the human collaborating with the robot in the same workspace. Six works have been retrieved that applied primarily ANNs, embodied intelligence, and fuzzy control techniques for this type of CI task and interaction (Table 2).

Table 2. Description of the retrieved articles that use AI for CI tasks related to robot control programming, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Robot control programming	[52]	Robot hybrid force and position control	Control	Adaptive ANN	Machine learning	NS
	[53]	Robot impedance control	Control	ANN and fuzzy impedance controller	Machine Learning, Logic-based	Industrial environment
	[54]	Robot motion control and adaptation	Control	DDPG and ANN	Embodied Intelligence, Machine learning	Manufacturing industry
	[55]	Shared control arbitration strategy for teleoperation	Control	DDPG with decision point reward term	Embodied Intelligence	NS
	[56]	Robot speed control and adaptation	Perception, Control	Multi-modal sensor fusion algorithm and CNN for human tracking and fuzzy inference system for robot speed control	Machine learning, Logic-based	Manufacturing industry (aviation)
	[57]	Shared control method for cooperative driving	Control	Approximate dynamic programming with neural networks	Machine Learning, Optimization	Automotive industry

The acronyms used have the following expansions: Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Deep Deterministic Policy Gradient (DDPG), and Not Specified (NS).

The selected papers categorized under robot control addressed CI tasks such as the learning of force and position control [52], impedance control [53], robot motion and speed adaptation for safety [54,56], shared-control strategies between the operator's inputs and the autonomous system's inputs for teleoperation [55] and driving [57]. Specifically, the solutions were developed for manufacturing picking, lifting and assembly collaborative

tasks, for which the effective control of cooperation and conflict between machines/robots and human operators is critical for safety.

Hybrid force and position control is required in collaborative scenarios where the robot can be in contact with humans, and both position and force trajectories need to be tracked, or alternatively an impedance controller can be used for compliant robot behavior when contact with a human is expected [53], such as for exoskeletons, assisted lifting for heavy industrial tasks, or learning from demonstration that requires direct manipulation of the robot. Low tracking error, stable control, the handling of uncertainties in the robot's dynamic model, and the optimization of human comfort are design requirements for safe human–robot interactions [52]. Robot speed scaling can be further applied according to the estimated risk and hazard level of collision. As previously mentioned, for cooperative driving tasks, the aim is to reduce the driver's workload, ensure safety, and improve the commands' efficiency compared to direct driving/teleoperation, requiring intelligent control solutions [55,57].

Among the selected articles, the majority of contributions were categorized within the machine learning, logic-based and embodied intelligence paradigms for control problems. For effective and adaptive intelligent control, we highlight the use of neural network-based or fuzzy controllers. Neural networks can estimate complex dynamic models of a robot and adapt to changing environments, while fuzzy controllers allow to account for the subjective or imprecise target variables, such as the human effort in a task or speed scaling based on a collision risk assessment.

Ref. [53] developed a cooperative impedance fuzzy-controller to optimize human–robot cooperation in heavy load lifting tasks, by mapping the interaction force, the derivative of the interaction force, and the end-effector velocity to a fuzzy impedance assistance level, followed by the use of a feedforward neural network to select the best assistance level that minimizes human effort/physical stress and maintains trajectory smoothness. A fuzzy control strategy can also be applied for robot speed control to ensure the safety of human–robot interactions in manufacturing work cells, while minimizing robot stops [56]. A fuzzy-based control strategy was developed to monitor the time derivative of the distance between the operator and robot, the velocities, and the temperature of the closest point to the robot (to distinguish between human and non-human surfaces), and to output a velocity scaling factor based on five rules, mapping the inputs to three fuzzy sets of risk level (high, medium, and low). Alternatively, the motion of the robot can be adapted when a human is too close, changing its path but not the velocity [54]. For this control problem, a reinforcement learning method can be trained offline in a simulated environment to learn when the collision risk is too high and how to change the robot's motion to the target point by adding a collision cost to the RL model's reward function.

An RL reward function can also be designed to reward the allocation of control authority to a human when the robot has multiple paths/goal options, allowing one to limit the search space, and ensuring assistance to the user's intended goal by exploiting the already existing information on points of disagreement between sub-policies (when the agent cannot predict the user's goal). Ref. [55] proposed such an RL framework for learning a shared control arbitration strategy for teleoperation applications, where the agent policy is modeled as a mixture model over each sub-policy towards a goal, therefore allowing for the identification of the decision points (multiple goal options) by the modality of the mixture distribution. The results in a simulated environment showed that the learned control strategy for blending robot and user control was safer and more effective compared to direct teleoperation by biased and noisy users.

For safe human–robot collaboration, several controller solutions were proposed for adaptation of the robot behavior according to the dynamic risk of hazards to the human workers. The collision risk can be estimated from input human and robot position and velocity data, while the workers' physical stress/effort can be retrieved from the interaction force with a cobot or from electromyography sensor signals (electric activity of muscles) for mobile monitoring [53]. This mapping between the task data and risk level can be

implemented by simple neural networks or simple safety rules. In addition to the preventative measures, collision detection and mitigation solutions can reduce the hazard to human operators, if the collision avoidance methods fail. Impedance control and human recognition by the detection of the collision surface's temperature can be employed together to limit the robot's collision force when necessary [56].

Shared control strategies can improve the safety and efficiency of control, while still respecting the operator's intent. However, the lack of explanation or transparency of the system's autonomous actions can affect the situational awareness of the human, and cause overtrust or distrust towards the machine [5]. More research is needed to determine the best way to communicate control takeover and control actions, and how the communication method affects the subsequent human behavior. Ref. [58] reviewed several methods for communication between the robotic system and human operator in shared control scenarios, mostly focused on feedback modalities for communicating environmental information or assistive cues that indicate the desired operator action, relayed through the master manipulator. Still, visual, haptic, or tactile modalities can also be employed to communicate the robotic system's intent to the human in the least sensory-taxing way.

In some of the reviewed works, the learning and testing of safety-critical interactions and functionalities were also conducted in simulated environments, before implementation in real collaborative tasks, such as the testing of position and force tracking for a collaborative cell [52], and the offline training of a Deep RL model and a hazard estimator with a simple humanoid model built from the recorded skeleton data of human motion [54]. A common safety requirement that is overlooked is the assessment of the quality and representativeness of the training data, followed by a thorough validation of the required performance levels and robustness in all operational conditions [5]. Typically, the focus of the papers is on the technological implementation, using the tasks as a way to evaluate the proposed solution, instead of validating it in the expected operational context, environment, and user type.

3.1.3. Design Optimization

Several requirements and their trade-offs need to be considered when determining the best design parameters for a particular application or task, i.e., design optimization. This can be challenging when the design space is large and high-dimensional, requiring the use of efficient algorithms to explore and evaluate a representative subset of designs. Additionally, it is commonly difficult to incorporate subjective factors into the optimization process, such as human preference or expert knowledge [59]. Only two works were retrieved applying an AI method for design optimization (Table 3) (regarding the AI problem domain classification in the aforementioned taxonomy, the defined categories fall short for tasks such as design optimization that can involve the challenges of multiple domains).

Table 3. Description of the retrieved articles that use AI for CI tasks related to design optimization, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Design optimization	[60]	Design optimization for wearable robot	ND	Bayesian optimization	Search and Optimization	NS
	[61]	Software design optimization	ND	Hybrid evolutionary and local search algorithm H-EDA	Search and Optimization	Rail industry

The acronyms used have the following expansions: Hybrid of the Estimation of Distribution Algorithm and an Iterated Local Search Algorithm (H-EDA), Not Defined (ND), Not Specified (NS), and reinforcement learning (RL).

Customization is another key goal of Industry 5.0. The results retrieved from our search categorized under Design Optimization CI tasks highlight this trend, addressing approaches that can help optimize the design of wearable robots, collaborative robots, human–machine interfaces and interaction devices, and software to the individual user preferences, characteristics, and requirements.

Ref. [60] proposed a human-in-the-loop design optimization approach for wearable robots (we do not consider this a situation where the human assists the machine, as it is not the expertise of the human that is used to optimize the design of the robot but force and kinematic data collected from specific movements), to minimize human metabolic energy during physical movement (computed using a musculoskeletal model of the human and the human–robot coupling). The approach demonstrated that it could reduce the total metabolic cost of walking by 24.91% using an ankle exoskeleton compared to an unassisted condition, while reducing the design time and cost of wearable robots. A similar work explored previously from [53] proposed to optimize a collaborative robot's impedance control parameters for assisted lifting tasks but using instead real-time feedback from the human's physiology. Both options can be employed for design optimization, depending on the design parameter of interest. Alternatively, design optimization can also be applied to software solutions. Ref. [61] proposed a safety software design method for safety-critical systems using Boolean algebra and optimization techniques to find an optimized software design that considers system performance, availability requirements, and safety rules. The proposed solution can facilitate the human-based design activity of safety-critical systems and can be adapted to other applications, domains, and standards.

Both articles reviewed fall under the Search and Optimization paradigm category, specifically using Bayesian optimization and a Hybrid evolutionary and local search algorithm. Bayesian optimization, commonly used in ML to optimize hyperparameters, effectively optimizes objective functions by building a surrogate model of the objective, quantifying the uncertainty by Gaussian process regression, and using the surrogate to define an acquisition function to choose samples to evaluate the real objective function. Using a case study on the design of an ankle exoskeleton [60], the optimization approach demonstrated the ability to optimize the spring stiffness design parameter, reducing the total metabolic cost of walking by 24.91% compared to an unassisted condition. The software design optimization solution was implemented and tested instead on a simplified railway case study, evaluating several optimization techniques based on the number of safety rules violated, the availability of the solution (in terms of the availability of green traffic lights for trains), and performance (evaluated in terms of the network traffic flow capacity). The multi-optimization results showed that all methods achieve a safe design quickly, but the best performing method is H-EDA, a hybrid of an Estimation of Distribution Algorithm (EDA—an evolutionary algorithm that samples new solutions at each iteration from a probability distribution) and an Iterated Local Search algorithm (ILS—an extension of a local search algorithm that iteratively searches for local minima). The hybrid algorithm uses the high-performing and problem-agnostic EDA algorithm as the main method and then applies the fast ILS to the best solution.

At the stage of system and design optimization, risk assessment and risk mitigation measures should be taken. Different methods can be used, such as task-oriented methods, automated formal verification or simulation techniques [62] that need to account for human error and unsafe human behavior. Human-centered design can promote performance optimization under safety constraints instead of considering safety measures as being limiting to performance [62]. Recent work has demonstrated the benefit of including cognitive ergonomics [63] and robot–human interaction quality metrics [64] for the design of CI systems.

3.1.4. Human/Object Detection and Tracking

Human or, more generally, object detection and tracking is a common task carried out by computer vision. From the retrieved papers, three employed intelligent vision systems to assist the human in collaborative tasks, while autonomously ensuring appropriate safety levels (eliminating the need of human supervision for safety monitoring). A fourth paper performed human tracking but with the goal of the real-time estimation of task advancement for task coordination between humans and robots (Table 4).

Table 4. Description of the retrieved articles that use AI for CI tasks related to human/object detection and tracking, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Human/object detection and tracking	[65]	Object tracking	Perception	GPLVM as generative prior and object flow classifier	Probabilistic Machine Learning	Manufacturing industry
	[66]	Human task tracking	Perception	Dynamic time warping-based algorithm	Search and Optimization	Manufacturing industry
	[67]	Human detection	Perception	4 CNN models (3 use transfer learning)	Machine learning	Industrial environment
	[56]	Human tracking	Perception, Control	Multi-modal sensor fusion algorithm and CNN for human tracking and fuzzy inference system for robot speed control	Machine learning, Logic-based	Manufacturing industry (aviation)

The acronyms used have the following expansions: Convolutional Neural Network (CNN), Gaussian Process Latent Variable Model (GPLVM).

Within complex industrial environments, perception is an essential ability for quality assurance, the detection of production deviations, and the improvement of productivity and safety in human–robot collaboration activities. Ref. [65] proposed a robust and universal object tracking and displacement estimation approach for industrial environments that can be applied to human workers or other obstacles of interest, aiming to tackle the common challenges of object tracking, such as unpredictable human motion, low discriminant information between targets and background, occlusions of target objects, and frequent change in the background or the targets' appearances (workers with casual clothing or Personal Protective Equipment (PPE)). Some of the proposed past works have limited applicability [65,67], due to the use of a single camera and simple classification labels, not allowing one to distinguish between blind spots or overlap situations from real collisions. The work of [56] proposed a solution to these issues by developing a multi-modal perception system, with two depth sensors to avoid occlusions and a thermal camera to reduce false positives.

Tracking can also be employed for task advancement monitoring in real-time [66] by tracking human motion and comparing the ongoing tracked movements with a task reference demonstration, providing an estimate of the task duration. These types of solutions are developed to deal with the variability inherent in human task execution, between subjects and within the execution by the same subject, which can be a challenge for collaboration and coordination tasks.

For these perception problems, a variety of AI paradigms and approaches were proposed, including Convolutional Neural Network (CNN)-based methods and transfer learning. Ref. [65] merged more traditional methods, such as object flow-based tracking and a particle filter tracker (a probabilistic approach for system state estimation from partial/noisy observations), to estimate the tracking target object's displacement and direction of movement from sequences of images, ignoring other movements in the scene. The object flow was tracked with a binary margin-based classifier, trained with positive and negative samples of object displacement. Additionally, a Gaussian Process Latent Variable Model (GPLVM; the GPLVM is a model that uses Gaussian processes to capture the latent structure of high-dimensional data in an unsupervised way and map it to a low-dimensional generative representation model of appearance) was used as a generative probabilistic prior to tracking the appearance changes of the human and consequently increasing the robustness of the approach. Refs. [56,67] both employed CNNs instead for the task of human detection in human–robot interaction scenarios. CNNs are particularly suited for image- and video-based classification tasks, and transfer learning can be employed with pre-trained networks to decrease training time, improve human–robot collision detection performance, and to take advantage of the human motion knowledge already existent in large open datasets [67].

For multi-modal detection and human tracking, ref. [56] proposed a technique to fuse the modalities (depth and thermal image information) into a single image that can be processed by the CNN. Human tracking was performed by first segmenting the static environment from the dynamic entities to be tracked, generating point-cloud clusters. Then, a sensor fusion algorithm called the RGB Mapping Approach (RGB-MA) was used to merge the multi-modal image information into a single image, mapping it into the RGB channels. Subsequently, a pre-trained YOLOv3 CNN was re-trained with a dataset of fused images to draw a bounding box around the area where the human was detected. The performance of the method was compared to a standard pre-trained RGB-CNN (for human detection on RGB images) and single-modality CNNs. The pre-trained RGB-CNN could not provide reliable detection results, confusing humans with other objects with similar shapes. Compared to the single-modality CNNs, the results showed that the multi-modal method could better distinguish humans from non-human objects due to the combination of temperature and spatial information, achieving the lowest percentage of false positives (2.4% versus 36.87% with only depth information and 64.35% with only temperature information), which the authors claimed can help to reduce unnecessary robot stops.

Human motion tracking for task advancement monitoring brings different challenges. The method proposed to deal with this is a modified version of the Dynamic Time Warping algorithm that allows one to align signals in time, measure their similarity, and deal with occlusions or human movements not relevant to the task [66]. The template of the task is learned online from previous executions not requiring offline training. The average execution speed can be computed from the estimated task advancement and used to estimate the remaining task duration. The method was tested in three realistic assembly operations. Compared to the elapsed time-based algorithm, which measures activity advancement through the elapsed time and past executions of the task, the proposed method achieved lower task time estimation errors. Future work to detect task execution errors and operational variants can greatly improve the robustness of the method.

Despite the benefits of some of the proposed strategies to human safety in industrial environments, security and privacy issues are being raised, as these systems should be resilient to cyberattacks and should keep sensitive and personal worker data safe from unauthorized access [68].

3.1.5. Human Intention Prediction and Motion Recognition

The type of CI task that was more frequently addressed by the retrieved research work was human intention prediction and/or human motion recognition, specifically for human–robot collaboration applications, in which the human is regarded as an uncontrollable autonomous agent. These two tasks have been addressed separately in some works, but as they are connected and commonly are tackled together (e.g., prediction of human motion intention), this section includes papers on both tasks. Twenty-five works have been retrieved that proposed AI solutions for this type of CI task and interaction (Table 5).

Within the two categories of human intention prediction and motion recognition, we highlight the variety of CI tasks that were addressed by the selected works. The amount of research focused on the investigation and development of solutions for these CI tasks corroborate the current lack of human-aware systems with the ability to understand the human co-worker's intentions and task objectives. Motion intention prediction and goal inference from motion are common challenges in HRC scenarios that can not only support safe collaboration from collision avoidance techniques and human–robot contact intention detection but can also be employed by cognitive systems for worker workload alleviation automation and assistance. In the automotive domain, human intent prediction is performed by Advanced Driver Assistance Systems (ADASs), automatic collision avoidance, and cooperative control systems. Similar tasks, such as gesture and motion recognition, are developed for HRC safety and human–robot teleoperation/instruction.

Table 5. Description of the retrieved articles that use AI for CI tasks related to human intention prediction and motion recognition, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Human intention prediction and motion recognition	[69]	Intention prediction	Perception	3 standard classifiers: SVM, RF, and LR	Machine learning	Automotive industry
	[70]	Gesture recognition	Perception	HMM	Probabilistic	Manufacturing industry
	[71]	Motion goal inference	Perception	Bayesian Inference of human goal	Probabilistic	NS
	[72]	Motion recognition	Perception	CNNs with transfer learning	Machine learning	Manufacturing industry
	[73]	Human activity intention prediction	Perception	Higher-order Markov chains	Probabilistic	Manufacturing industry
	[74]	Intention prediction	Perception	HMM and Probabilistic Dynamic Time Warping	Probabilistic	Automotive industry
	[75]	Gesture recognition	Perception	ANNs	Machine learning	Manufacturing industry
	[76]	Motion intention prediction	Perception	HMM	Probabilistic	Manufacturing industry
	[77]	Motion intention prediction	Perception	PDMP	Probabilistic	NS
	[78]	Motion intention prediction	Perception	Gaussian processes	Machine learning	Industrial environment
	[79]	Human behavior prediction	Perception, Control	3 standard classifiers (SVM, DT, LR)	Machine Learning, Probabilistic	Automotive industry
	[42]	Motion intention prediction	Perception, Planning	Probabilistic movement primitives, GMM	Probabilistic	NS
	[80]	Intention prediction	Perception, Planning	CNN	Machine learning	Aviation industry
	[81]	Action intention recognition	Perception	3D-CNN for HAR and 1D-CNN for contact detection	Machine learning	Manufacturing industry
	[82]	Action intention prediction	Perception, Planning	RF and XGB (Ensemble learning methods)	Machine learning	Aviation industry
	[83]	Motion intention prediction	Perception	LSTM-RNN	Machine learning	Industrial environment
	[84]	Motion intention prediction	Perception	LSTM-RNN	Machine learning	Manufacturing industry
	[85]	Human behavior prediction and preference adaptation	Perception, Optimization	Hierarchical RL	Embodied Intelligence	Automotive industry
	[86]	Motion recognition	Perception	Transfer learning with CNN	Machine Learning	Manufacturing industry
	[87]	Action recognition	Perception	Transfer learning with ST-GCN	Machine learning	Manufacturing industry (aviation)
	[88]	Intention recognition	Perception	LSTM-RNN	Machine Learning	NS
	[89]	Intent recognition	Perception, Control	CCA and OTWV to classify EEG signals and laser obstacle detection system	Embodied Intelligence	Automotive industry
	[90]	Action recognition	Perception	LSTMs	Machine Learning	NS
	[91]	Error recognition	Perception	SVM to classify EEG signals	Machine Learning	Manufacturing industry
	[92]	Intent recognition	Perception	DRNN	Machine Learning	Manufacturing industry

The acronyms used have the following expansions: Artificial Neural Network (ANN), Canonical Correlation Analysis method (CCA), Convolutional Neural Network (CNN), Decision Tree (DT), Deep Residual Neural Network (DRNN), Gaussian Mixture Model (GMM), Electroencephalography (EEG), Hidden Markov Model (HMM), Human Action Recognition (HAR), Logistic Regression (LR), Long Short-Term Memory Recurrent Neural Network (LSTM-RNN), Not Specified (NS), Overlap Time Windows Voting method (OTWV), Probabilistic Dynamic Movement Primitive (PDMP), Random Forest (RF), Spatial–Temporal Graph Convolutional Networks (ST-GCNs), Support Vector Machines (SVMs), XGBoost-Extreme Gradient Boosting (XGB).

In human–robot collaboration applications, human motion recognition is performed using CNN models combined with transfer learning [72,86,87] for improved recognition performance. Another option is the use of skeleton data, where the relationships between the joints provide information about the worker's motion/activity. Ref. [90] trained an LSTM model with spatiotemporal human-skeleton data for activity recognition, achieving 91.4% accuracy on a dataset.

To increase human–robot collaboration efficiency, a number of works focused on predicting the human motion intention. The most commonly used models for this task were LSTM-RNN networks, when using physiological signals' patterns (e.g., EEG, myography, and eye-tracking) [83,84,88], Multivariate Gaussian Distribution-based ML models [78],

and Markov Chain-based methods for modeling human intention based on sequences of human activities or postures [73,76]. Human motion intention recognition may be context-dependent, as similar motion may correspond to different tasks. To better differentiate between similar situations, ref. [72] proposed an extra tool identification step after motion recognition, implemented by another deep neural network, to better understand the task context. The results from training and testing with motion images of 90 human subjects showed that the model was able to achieve a human motion classification accuracy of 96.6%, and was able to identify the tools that the human was holding. For HRC scenarios that require physical human–robot contact, vision and tactile perception data can be combined for interaction safety monitoring. Ref. [81] proposed the use of two deep CNNs to combine the modalities, achieving the highest mean accuracy by the visual recognition system of 99.7%, while for the contact detection system, the highest mean accuracy was 99%, being also able to detect the type of contact (intentional or incidental).

Human gestures are also useful for more intuitive and efficient robot control and instruction, providing a more ergonomic interaction modality, compared to traditional teleoperation systems and controllers. Ref. [75] developed a gesture-based human–robot interaction framework that allows the worker to give instructions to a collaborative robot for tools and parts' delivery, or for holding objects. The human's upper body gestures were captured through wearable inertial measurement units, and with an unsupervised sliding window technique, the data were segmented into static and dynamic sections. Then, multiple feedforward ANNs were used to classify the different types of dynamic and static gestures. Additionally, a parameterization robotic task manager algorithm was implemented to translate the recognized gestures into robot commands, and provide visual or speech feedback to communicate to the human worker: the gesture options available to select, if the gesture was recognized, and what gesture was identified. The proposed solution was evaluated in a manufacturing assembly task, demonstrating that it is effective in recognizing in real-time 12 different types of gestures in a subject-independent manner, achieving accuracies between 90% and 100%.

A trend we identified in the reviewed body of work related to safe HRC systems was the use of human physiological signals for motion intention recognition, employing deep Long Short-Term Memory (LSTM)–Recurrent Neural Network (RNN) networks for dynamic time-series processing. Brain waves monitored with mobile EEG systems [84], body movement data collected with force myography bands [83], and human gaze data from eye-tracking and VR headsets have been shown to be able to detect the worker's intentions from 54 milliseconds up to 2 seconds before the action, providing the robot with time to take initiative. EEG has also been employed to detect interaction errors/conflicts between the worker's intention and the robot's actions, which can be employed to improve the system's response to errors [91]. The evoked error-related potentials in the brain, due to either a robot error in the user intention detection, or violation of the user's preference during autonomous behavior, were used by an SVM model to correctly distinguish the type of error above the chance level.

On another hand, the worker's interactions can be classified indirectly from the robot sensor data, and the operational context can be classified from the interaction duration, as well as if the interaction was soft or hard, with what part of the robot it took place (tool or link), and if it was a planned interaction or accidental. To achieve the soft object/hard object interaction classification from robot torque measurements, ref. [92] employed a widely used neural network, the ResNet, achieving 98% accuracy on the test dataset.

The semi-autonomous and autonomous driving field in particular has brought to the AI community several advancements stemming from the complex research problems it needs to tackle, namely, perception, probabilistic modeling, human–machine interaction, and multi-agent decision-making [93]. For these applications, ML methods, such as SVMs [69,79], Ensemble Classifiers [69,82], and HMMs [74] are the most commonly used AI methods for the driver's intention and behavior prediction, achieving prediction accuracies above 90%, while only one selected work used a Hierarchical Reinforcement

Learning approach for learning the driver's preferences and adapted an advanced driver assistance system feature to it [85]. Probabilistic methods can further be incorporated to account for the uncertainty inherent to human behavior. An example is the work of [79] that improved the prediction accuracy of standard classifiers by including the worst possible scenario safety information derived from Hamilton–Jacobi (HJ) reachability. The drivers' behavior was predicted with a probability between 0.75 and 1. Another alternative is the direct observation of user gaze behavior and the probabilistic modeling of the eye data for driver intention prediction. Ref. [74] modeled the observed eye-gaze/eye-fixation data temporal patterns with Probabilistic Dynamic Time Warping (DTW) distributions, and the underlying human intention as the latent states of a Hidden Markov model (HMM) to predict the driver's maneuver intention. The novel extension of the DTW method to be probabilistic and using it to capture the similarities between the observation patterns corresponding to different intentions allowed the framework to anticipate in real-time the drivers' intentions approximately three seconds before the maneuver, with more than 90% accuracy. Reaching good robustness and generalization to new human subjects is a common issue with systems that aim to model human behavior. The work of [85] was able to capture the driver's preferences and variability with a Hierarchical Reinforcement Learning model, for an adaptive and fairness-aware personalized ADAS. The method involves using three interacting RL agents to learn to adapt to intra-, inter-and multi-human state variability, applying Q-learning-based algorithms. The framework was evaluated in a simulated environment and the results for simulated human subjects with different driving behaviors showed that the performance of a time-to-crash alarm was adapted iteratively to different subject behaviors.

Taking into account a similar type of task and automation goal but applied to the aviation and air control domain, ref. [80] used a CNN for the prediction of controller actions/strategy to develop a personalized workload alleviation automation solution. The method processed the visual features of air traffic scenarios (represented by a velocity obstacle diagram that provides all the necessary information for informed conflict resolution) and achieved significantly better results with the personalized models compared to general models (non-individualized). Also, ref. [82] aimed at reducing the controllers' mental workload by predicting and automating some tasks. Here, ensemble ML methods, namely, Random Forest and XGBoost-Extreme Gradient Boosting, were used to predict the workers' actions using the aircraft's trajectory features. The models for altitude, speed, and course change prediction all achieved accuracies over 80%, with the highest of 99% for vertical maneuvers.

Within this category of CI systems, the safety measures implemented are mainly focused on the assistance of human workers by reducing the workload and by safety monitoring (as in the case of autonomous collision avoidance). In the case of direct human–robot contact, such as in collaborative tasks and learning from demonstration scenarios, the recognition of accidental or intentional interaction may be applied for hazard mitigation. A third contact intention category has not been considered in the reviewed works, but it may also have safety and security implications: intentional interaction with a robot can be further categorized as assistive (positive intention) or obstructive (negative intention), depending on the intention of the worker to assist with the current task or hinder it. To resolve this intention conflict, methods similar to the ones used for shared-control arbitration may be appropriate (see the related works reviewed in Section 3.1.2).

3.1.6. Human State Recognition

Research works addressing the task of human state recognition vary in the target state to be identified, depending on the final application domain. For human–robot collaboration, the human physical state might be of higher importance when the goal is to reduce physical human effort. In less physical safety-critical applications, such as driving or piloting or for operators of control rooms or air traffic controllers, the mental, cognitive and affective states of the user tend to have greater importance and impact on human error. Ten works

have been retrieved that proposed AI solutions for this type of CI task and interaction (Table 6).

Table 6. Description of the retrieved articles that use AI for CI tasks related to human state recognition, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Human state recognition	[94]	Human state recognition	Perception	DBSCAN	Machine learning	Automotive industry
	[95]	Human drowsiness recognition	Perception	Modified hierarchical PSO-H-ELM	Search and Optimization, Machine learning	Automotive industry
	[96]	Human impedance estimation for exoskeletons	Perception	RF	Machine learning	NS
	[97]	Human performance prediction	Perception	SVM with bootstrap aggregation	Machine learning	Nuclear industry
	[98]	Human emotion recognition	Perception	Transfer learning using deep CNN	Machine Learning	Manufacturing industry
	[99]	Human mental workload and performance prediction	Perception	Multiple linear regression	Machine Learning	Aviation industry
	[100]	Human attention state recognition	Perception	Multiple linear regression DCNN and PCA for feature fusion, and selection kNN and SVM for classification	Machine Learning	Automotive industry
	[101]	Human emotion recognition	Perception	Hybrid model with DNNs and SVMs	Machine Learning	Automotive industry
	[102]	Human mental workload recognition	Perception	CNNs for feature extraction, LSTM for classification	Machine Learning	Industrial Environment
	[103]	Human behavior and characteristics estimation and adaptation	Perception	A-POMDP model for behaviour prediction, ABPS for robot policy adaptation	Probabilistic	Industrial Environment

The acronyms used have the following expansions: Adaptive Bayesian Policy Selection (ABPS), Anticipatory Partially Observable Markov Decision Process (A-POMDP), Convolutional Neural Network (CNN), Deep Convolutional Neural Network (DCNN), Deep Neural Network (DNN), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Extreme Learning Machine Algorithm with Particle Swarm Optimization (PSO-H-ELM), K-Nearest Neighbor (kNN), Long Short-term Memory model (LSTM), Not Specified (NS), Principal Component Analysis (PCA), Random Forests (RF), Support Vector Machines (SVM).

Human state recognition has been proposed by the works selected for four main applications: for driver state recognition, namely, drowsiness, attention and affective state, for mental workload recognition in aviation, for human performance prediction in control rooms, and for physical and affective state recognition in HRC scenarios. These tasks can be classified as perception tasks directed towards the awareness of the workers' internal state, from physiological or behavioral indicators of state. Different ML techniques can be leveraged for state classification, regression, or clustering tasks, from ensemble SVMs or random forest models, to deep learning models and unsupervised clustering. Several methods achieved high recognition accuracies (over 90%), indicating the potential of having human-state aware adaptive systems in the near future. The work of [103] took the first steps to accomplish complete human adaptation, including worker behavior prediction and adaptation on a short-term, and long-term policy adaptation to operator characteristics, such as expertise.

Four papers have proposed AI solutions for driver state detection, including (a) density-based spatial clustering of applications with noise (DBSCAN) for unsupervised clustering of the physiological data of drivers (electrocardiography and respiration signals) into clusters of driver behavior event, noise or normal behavior [94]; (b) Modified Hierarchical Extreme Learning Machine algorithm with Particle Swarm Optimization (PSO-H-ELM) for driver drowsiness detection from EEG data [95]; (c) a hybrid model for the fusion of deep learning-based and handcrafted image features, and classification of distracted driving with K-Nearest Neighbor and SVMs [100]; and (d) another hybrid model combining an SVM classifier with deep neural networks for the detection of drivers' emotions using video [101]. An average recall rate of 75% was achieved for driver abnormal event detection with real-world driving data; a mean accuracy of 83.67% was achieved for driver drowsiness detection with driving data collected from a simulator; the best accuracy of

95.9% was achieved for the detection of distracted driving using 500 features selected from a dataset with 10 classes; and the highest accuracy in facial emotion recognition was 98.64% for a dataset with six emotion classes. We highlight the use of Particle Swarm Optimization for the selection of optimal neural network hyperparameters, and the combination of the classical SVM classifiers with hand-crafted and deep-learning-based features for optimal recognition performance.

On the human–robot collaboration domain, ref. [98] aimed at recognizing human emotion for improved assistance and safety of the human partner. A standard web camera was used to monitor human emotion in real-time, and a transfer learning approach was used with a Deep CNN for improved classification accuracy. The highest cross-validation accuracy value reached was 97.82%. State recognition applications employing image or video data benefit from the large vision datasets available online; however, the limited practical robustness due to sensor constraints (worker pose variation or occlusion to the camera view), the biased and indirect state information that can be derived from behavioral cues (as opposed to physiological information, which cannot be controlled by the human and are directly affected by internal and external stimuli), and worker privacy laws are outstanding challenges.

In the control room application domain, ref. [97] aimed at predicting human performance using multi-source physiological data fusion. SVMs integrated with bootstrap aggregation were applied to fuse multi-modal physiological data (eye-tracking data, skin conductance response, and respiratory function) and learn to predict operator performance. The bootstrap aggregation was used to train an ensemble of 100 models with randomly selected samples, achieving accuracies between 75% and 83% on an independent dataset, outperforming the individual models.

A multi-modal physiological approach was also adopted by [99,102]. Ref. [99] employed EEG, eye activity, and heart rate variability (HRV) to assess the mental workload effects of changing task load in tracking and collision prediction tasks. With a dataset of 24 participants using multiple regression, 54.3% of the variance in the tracking task performance values could be explained by the model. For the collision prediction task, 61.7% of the variance in the results was explained by the model. With EEG, photoplethysmogram (PPG), and electrodermal activity (EDA) modalities instead, and using 2D and 1D CNNs for feature extraction and an LSTM model for binary workload level classification, an accuracy of 86% was achieved [102].

The work of [96] was focused on recognizing a physical human state, performing online estimation of human arm impedance/stiffness from surface electromyography and stretch sensor data, for the improvement and tuning of the exoskeletons' strength amplification controller. Using a random forest regression model, the estimation performance was evaluated offline with a validation dataset and online (no ground truth) by monitoring the controllers' stability while the arm stiffness was changing. The estimator achieved an online R factor of 0.993, outperforming the state-of-the-art results and improving the strength amplification bandwidth when compared to a robust controller that needs to have a conservative bound of human stiffness.

Human–machine communication and interaction performance can be impaired by the cognitive biases and expectations underlying the human subjective experience, further impacted by time pressure, fatigue, and stress [5]. The potential of human state recognition solutions for CI is clear; however, several challenges remain for industrial implementation, particularly for safety-critical systems. The human state and performance are task dependent and subject dependent, as workers can respond differently to different stimuli and task requirements, depending on their skills and initial state. This additional context is commonly disregarded by machine learning models and may be required for better generalization across domains. Moreover, for pervasive, continuous, and long-term state monitoring, the temporal dynamicity of most states and data distribution evolution should be taken into account [104].

3.1.7. Human Mental Model Estimation

A different type of human-related estimation was studied in the works of Tabrez et al. (2019) [105] and Blum et al. (2022) [106] (Table 7). Estimating the human mental model, meaning a human's mental representation of how a system works, provides insights into how operators perceive and interact with a system, supporting the development of improved user-centric designs, better communication in collaborative tasks, modeling of the decision-making process, detection of cognitive biases, prediction of human behavior and performance, and detection of potentially dangerous situations stemming from human-system model disparities.

Table 7. Description of the retrieved articles that use AI for CI tasks related to human mental model estimation, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Human mental model estimation	[105]	Infer Human's behavior reward function for human–robot joint tasks	Perception	HMMs	Probabilistic	NS
	[106]	Modelling of Pilot's mental models for behaviour anticipation	Knowledge representation, Reasoning, Perception	HCA for behaviour pattern identification and ACT-R cognitive architecture for cognitive modelling	Logic-based, Machine Learning	Aviation industry

The acronyms used have the following expansions: Adaptive Control of Thought-Rational (ACT-R), Agglomerative Hierarchical Cluster Analysis (HCA), Hidden Markov Model (HMM), Not Specified (NS).

Ref. [105] addressed this concept, with a similar idea, in part, to the work of Kulkarni et al. (2019) [41] (in Section 3.1.1), for the development of a Reward Augmentation and Repair through Explanation (RARE) framework, aimed at improving human–robot collaboration by detecting model disparities between the robot and the human, identifying the source of the disparities, and providing explanation-based feedback to the human for the repair of the policy. Using a HMM to model the human mental model of the reward function, the viability and effectiveness of the method were tested with a joint-execution collaborative game with a robot, leading to more successful games and a more positive user experience than the control condition.

Ref. [106] used a cognitive computational modeling approach to simulate pilots' mental models and anticipate their behavior in human–AI teams. They first employed the unsupervised machine learning method Agglomerative Hierarchical Cluster Analysis (HCA) to learn and identify individual behavior patterns in situation models from complex and unstructured empirical data, also helping to identify individual differences in mental representations. The data used for the clustering analysis were based on the quantitative response times, log data, and visual focus measured with eye-tracking, from a flight simulator study with 13 participants. The cognitive architecture Adaptive Control of Thought-Rational (ACT-R) was then used to simulate situation models, based on data collected from pilot interviews. The models were implemented to mentally simulate different possible outcomes of action decisions and the timing of a pilot. Model tracing, which monitors the pilot's interactions with the system, allows for a comparison of simulated normative behavior and the pilot's individual differences, learned from sub-symbolic activation of the symbolic models' structures. The individually simulating model (ISM) performed significantly better (37% increase in accuracy in anticipating a first pilot action and about 13% for a second action) than the normative model (NM) in anticipating the pilot's behavior in an engine fire event. The group means for the anticipated reaction time showed a significantly smaller deviation with the ISM compared to the NM. It was demonstrated how the combination of ACT-R and machine learning can improve pilot assistance by selecting sub-models and learning from experience.

Safety measures should take into account how the workers mental model evolves over long-term interactions and how the created explanations for policy repair should adapt to the worker's confidence on their mental model.

3.1.8. Ci Safety Assessment

This section includes a variety of papers that can be included under CI for safety assessment, as they propose AI methods specifically developed to monitor or assess the behavior of a system, essential to guarantee safety in safety-critical applications. Eleven works have been retrieved that proposed AI solutions for this type of CI task and interaction (Table 8).

Table 8. Description of the retrieved articles that use AI for CI tasks related to safety assessment, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
CI safety assessment	[107]	Robot–human cooperation supervisor learning	Control task	L* learning algorithm for supervisor synthesis	Probabilistic, Logic-based	Manufacturing industry example
	[108]	Monitoring system for HRC workspaces	Perception	Faster R-CNN ResNet 101 Coco for object detection, ANN for safety state assessment and CNN for speech recognition	Machine learning	Manufacturing industry
	[109]	Human motion prediction uncertainty estimation for HRC	Reasoning, Perception, Planning	Bayesian framework for human motion prediction uncertainty reasoning	Probabilistic	NS
	[110]	Inference of model confidence of worker's movement prediction	Reasoning, Perception, Planning	Bayesian framework for human motion prediction uncertainty reasoning	Probabilistic	NS
	[111]	Driver assistance by a situation-aware advanced driver assistance system	Control	POMDP to model the driver and vehicle state and learning L* learning based algorithm to iteratively learn the supervisor	Probabilistic, Logic-based	Automotive industry
	[112]	Optimization of human–robot dynamic safety zones and robot stop trajectories	Planning, Control	Constrained non-linear optimization	Optimization	Manufacturing industry
	[113]	Software architecture for human–AI teaming in smart manufacturing	Perception, Reasoning, Control, Communication, Knowledge representation	Knowledge graphs and relational machine learning	Knowledge-based, Machine Learning	Manufacturing industry
	[114]	Explainable Multi-Agent Path Finding	Planning	XG-CBS	Search and Optimization	NS
	[115]	Invariant mining and validation for model checking	Perception, Knowledge representation	RL	Machine Learning	Rail industry
	[116]	Framework for Situational assessment, resource optimization and decision-making for anxiety mitigation	Perception, Reasoning	Mixed-integer programming and Gurobi optimizer	Logic-based, Optimization	Industrial environment
	[117]	Operator risk perception prediction	Perception	HSMM	Probabilistic	Manufacturing industry

The acronyms used have the following expansions: Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Explanation-Guided Conflict-Based Search (XG-CBS), Hidden Semi-Markov Model (HSMM), Human Robot Collaboration (HRC), Not Specified (NS), Partially Observable Markov Decision Process (POMDP), Region-based CNN (R-CNN).

Four general CI tasks were identified in the selected works focused on implementing safety functionalities: control tasks for safety and task requirement compliance, perception tasks for safety monitoring, model checking, and model prediction uncertainty reasoning.

Ref. [108] proposed a complete safety monitoring system for human–robot collaboration workspaces, using neural networks. Four neural networks were used for each component of the system: an object detector using the Faster RCNN ResNet 101 Coco model, safety distance assessors using ANNs, and a speech recognizer for communication between

humans and robots using a CNN. The object detector reached test set accuracies above 95%, the two ANNs used for human–robot safety distance assessment reached accuracies above 98%, and the CNN used as a speech recognizer in a preliminary test reached 93% accuracy in the classification of simple answers. In addition, for smart manufacturing applications, knowledge graphs can be leveraged to represent the expert knowledge associated with the product, manufacturing process, and operational data, augmented by knowledge derived from relational machine learning [113].

More specific supervisor models can be developed, such as the work of [107] that proposed the use of a L * learning algorithm [118] to ensure safety requirements and task completion in systems with uncertainties, such as human–robot collaboration. Here, the robot is modeled by Markov Decision Processes, the human by Partially Observable Markov Decision Processes (to account for their unknown/hidden intents), and the supervisor controller of the robot by a Deterministic Finite Automata. Probabilistic Computation Tree Logic (PCTL) is used to interpret the logic specifications of the system, and then a deterministic supervisor that satisfies the PCTL specifications can be learned by a L* learning algorithm, through queries and counterexamples. An example was given of a supervisor synthesized for a collaborative assembly task. The supervised assistance of the robot was able to help the human complete a task with fewer steps and a higher probability of reaching the final state. The method was extended in the work of Wu et al. (2021) [111] to formally demonstrate the convergence, soundness, and completeness of the method, and it was applied to another human–robot collaboration scenario: a driver and a semi-autonomous driver assistance system used to prevent steering off lanes. To prevent possible collisions between human and collaborative robots, ref. [112] developed a supervisor controller that focuses on optimizing stop trajectories. The method consists of the online scaling of dynamic safety bounding volumes enclosing the robot and human. The size of the zones is optimized by minimizing the time of potential stop trajectories, considering the robot dynamics, its torque constraints, and its speed with respect to the human. The optimization problem in this paper is solved using a open-source tool for non-linear optimization. Novel fluency metrics were proposed to evaluate and compare the approach to other existing methods that select a smooth stop trajectory among n tentative stop times, or that implement a static safety zone. The proposed method led to smaller safety zones and lower total task time.

Alternatively, ref. [115] proposed visualizations to assist engineers of railway signaling systems in model checking, by aiding in invariant mining. Reinforcement learning was employed in a first step, through simulation, to maximize the unique state observations of a Markov Decision Process and generate a dataset for invariant mining. The proposed visualizations aimed to increase the readability of the suggested invariants, which can be used to constrain the model checking to those states that are reachable.

The previous works proposed AI assistive solutions for improved human safety; however, a significant issue lies in the oversight of the quality and accuracy of the AI solutions, often excluded from risk assessments. Techniques can be developed to estimate the confidence/uncertainty on their outputs and communicate it to the human stakeholders. For human–robot collaboration, the preliminary work of [109] and the more recent one of [110] used a Bayesian framework for human motion prediction uncertainty reasoning, updating a Bayesian belief over the model's prediction confidence, modeled as a hidden state of an HMM, every time the human moved in a way that was assigned low probability by the model. Uncertainty quantification can also benefit the transparency of the algorithms [5].

Another option for robot oversight is the method of [114], an algorithm called Explanation-Guided CBS (XG-CBS) developed to balance planning time and explainability in safety-critical applications involving AI and humans. The XG-CBS algorithm modifies the Conflict-Based Search (CBS) method for Multi-Agent Path Finding (MAPF) by adding explainability constraints, resulting in a set of non-colliding paths that admit a short-enough explanation that can be visualized as a sequence of images for each time segment where the agent trajectories are disjoint. The results show that the approach is effective in balancing planning

time and explainability for the human supervisor. However, it may not always find the optimal solution and may require more planning time for larger environments.

We observed that more recent work displays a higher level of complexity [116,117], considering the several components of a HRC system that impact safety, requiring modules for situational assessment, resource optimization (e.g., dynamic task allocation), and decision-making for long-term risk mitigation. Notably, both works consider the worker's perception of risk as an important indicator for the application of mitigation measures.

3.1.9. Accident/Error Prediction

Accident/error prediction is an essential CI task, in which the AI predictive power can be used to prevent human-related critical situations in real-time. From the collection of retrieved papers, two applied accident and error prediction for driving applications, and the other two for operators in safety-critical industries (Table 9).

Table 9. Description of the retrieved articles that used AI for CI tasks related to accident/error prediction, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Machine Assists Human						
Accident/error prediction	[119]	Risk and accident prediction for a human-in-the-loop decision-support system	Perception	Bayesian network	Probabilistic	Maritime industry
	[120]	Adaptive forward collision prediction and warning for an ADAS	Perception	RL	Embodied Intelligence	Automotive industry
	[121]	Detection of operator error precursors for accident event prevention	Perception	Deep seq2seq neural network with attention mechanism	Machine learning	Nuclear and aviation industry
	[122]	Prediction of driver's speed tracking errors for an ASA	Perception	NAR-NN	Machine Learning	Automotive industry

The acronyms used have the following expansions: Advanced Driver Assistance System (ADAS), Advisory Speed Assistance System (ASA), Non-linear Autoregressive Neural Network (NAR-NN), Reinforcement Learning (RL).

The human aspect of accidents is still an outstanding problem that CI solutions can help with. Human assistance systems were proposed in the collection of selected articles, employing a variety of features that account for human factors or adapt to human preferences and behavior.

Elmalaki et al. (2018) [120] used Reinforcement Learning for adaptive forward collision prediction and warning, taking into account the driver's context, meaning the human state (specifically attentive or distracted state based on audio streams collected by the driver's phone) and preferences across drivers and time. The proposed driver-in-the-loop context-aware ADAS system led to an increase of 94.28% in the safety of the driver and a 20.97% improvement in the driving experience. A similar work [122] predicted the driver's speed tracking errors instead, using a Non-linear Autoregressive (NAR) Neural Network trained with data from N clusters of different types of drivers (different driving behavior). The developed Advisory Speed Assistance System (ASA) led to a reduction of 53% in speed error variance in simulated driving and a reduction of 20.00% of speed error variance in real vehicle experiments.

Other data such as statistical information about past accidents, or synthetic data emulating common human-system interactions can be used for the early detection and prevention of accident sequences. In the work of [119], a Bayesian network was applied for risk and accident prediction in ship navigation. The approach was developed to reliably predict accidents based on past data and statistical information about accident probabilities. A Bayesian network was used to model and reason about these data to provide future risk predictions. In [121], operator error precursors were detected for accident prevention in industrial human–machine systems. A common natural language processing deep-learning model, the seq2seq encoder–decoder model, was used to process visual control panel states from synthetic HMI data to detect operator error precursors. The evaluation of the model

in an out-of-sample dataset reached relaxed accuracies above 80%, and 10% difference from the actual HMI state sequences for the majority of the test cases.

3.2. Human-in-the-Loop Assists Machine

In this section, we summarize research works employing human-in-the-loop frameworks to assist the machine. After analyzing the collection of papers, a classification scheme based on the stages of an AI model lifecycle was used to group them: works where the human provides assistance during the learning process, or during the deployment process of the system. During the learning stage, the AI model is developed and trained in a training environment, and it is typically not integrated into the current system. In the deployment stage, the AI solution is moved to a production environment and integrated in the system that will be employed by the user.

3.2.1. Human Assistance During Learning

Most works involving human-in-the-loop assistance to the machine/intelligent system apply it during the learning process (Table 10), also called human-in-the-loop learning. Human feedback, particularly through demonstrations, can improve the efficiency of the learning process in terms of the number of samples required, compared to classical reinforcement learning, and be safer due to the interventions and corrective actions that the humans can provide in real-time [123].

Safe Training of RL Agents

RL-based systems that are designed to interact with humans can perform potentially dangerous actions during training. In particular, for model-free agents that can only learn with trial and error, having human intervention is the only way to avoid catastrophes when the agent has not learned yet. In [124], an RL agent was trained with a human preventing catastrophic actions of different types and complexity. In the scheme, not only was the unsafe action blocked but the human supervisor instead sent a safe action to the environment and returned to the agent a penalty for choosing the unsafe action. A CNN was then trained with state–action pairs and a corresponding binary label of whether the action was blocked or not, to learn to imitate the human's intervention decisions (perception of human policy) since human oversight for the total training time of an agent can be infeasible or very costly. Compared to a baseline RL approach with large negative rewards applied to catastrophic actions (which can suffer catastrophic forgetting), the human oversight RL approach performed much better at reducing catastrophes. The scheme is agnostic to the type of RL method, so it can be applied to different types of RL and to multiple agents.

Personalized Autonomous Driving

A human observer in the loop can also be beneficial for the optimization of reward models towards personalized assistance, such as for personalized autonomous driving behavior. The Progressive Optimized Reward Function (PORF) learning model proposed by [125] can be integrated into a Deep RL framework. A hybrid DNN model structure composed of a CNN followed by a conv-LSTM receives as input vehicle front-view sequential images labeled based on the vehicle driving behavior evaluation (the reward). The DNN is trained in two stages: in the pre-training stage, the network is trained with images labeled by a formula that represents the safety of the vehicle–road relationship using data from a virtual environment, and in the progressive optimization stage, the network is trained and optimized continuously with data from human evaluations of the vehicle driving behavior in a real environment. The PORF model showed an upward trend in cumulative reward with increased sampling and training, demonstrating the potential for human–machine collaboration in autonomous driving; however, the model's confidence should be measured before using the framework in new environments.

Table 10. Description of the retrieved articles that use AI for CI tasks where the human assists during the learning phase of intelligent systems, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Human Assists Machine						
Human assistance during learning	[126]	Learning from demonstration reactive and proactive robot controllers	Planning, Perception	Learning from demonstration with robot and human behaviour encoded in an ADHSMM	Probabilistic	NS
	[127]	Evaluation–feedback loop for interactive and iterative model generation	Reasoning	Bayesian network	Probabilistic	Industrial environment
	[128]	Learning robot trajectory from demonstration	Planning	Learning from demonstration with robot motion is encoded in a HSMM	Probabilistic, Embodied Intelligence	NS
	[124]	Human intervention for safe reinforcement learning	Perception	HIRL with CNN as human imitator	Machine Learning	NS
	[129]	Robot learning from demonstrations and subgoals	Planning, Perception	Human interactive IRL	Machine Learning	NS
	[130]	Human motion and coordination learning	Perception, Planning	Deep RL	Machine Learning, Embodied Intelligence	NS
	[131]	Interaction imitation learning of a robot policy	Perception	Imitation learning—extended DAGGER algorithm	Machine Learning	Automotive industry
	[132]	Language-based assistance for mobile agent vision-based navigation	Perception Communication, Planning	Imitation learning	Machine Learning, Embodied Intelligence	NS
	[133]	Learning and modeling of a dynamic system properties from demonstrations	Perception, Planning	VAR-POMDP model learned from human demonstrations using the Bayesian non-parametric learning method	Probabilistic	Automotive industry example
	[134]	Robot motion learning from demonstrations	Planning	GMR	Probabilistic, Embodied Intelligence	Manufacturing industry
	[135]	Learning from demonstrations for shared control strategy	Perception	Deep DDPG and online learning from demonstration	Machine Learning, Embodied Intelligence	Industrial environment
	[125]	Progressive learning of personalized autonomous driving	Perception	PORF learning model using a DNN integrated into a DDPG framework	Machine Learning, Embodied Intelligence	Automotive industry
	[136]	Learning variable impedance control human policy and reward	Perception	IRL from expert demonstrations	Embodied Intelligence, Machine Learning	NS

The acronyms used have the following expansions: Adaptive Duration Hidden Semi-Markov Model (ADHSMM), Convolutional Neural Network (CNN), Dataset Aggregation algorithm (DAGGER), Deep Deterministic Policy Gradient (DDPG), Deep Neural Network (DNN), Gaussian Mixture Regression (GMR), Hidden Semi-Markov Model (HSMM), Human Intervention Reinforcement Learning (HIRL), Human Robot Collaboration (HRC), Inverse Reinforcement Learning (IRL), Not Specified (NS), Progressive Optimized Reward Function (PORF), Reinforcement Learning (RL), Vector Autoregressive POMDP (VAR-POMDP).

Human-in-the-loop frameworks can also be employed for the iterative optimization of other AI models.

Interactive Alarm Flood Reduction

In [127], human assistance was used for an interactive alarm flood reduction task in a control room scenario by providing inputs in an evaluation–feedback loop of a probabilistic graphical model. In an alarm flood situation (defined as more than 10 errors per minute), data from error messages and warnings (alarm log) are employed as input for a probabilistic graphical model (Bayesian network), having used first a Max-Min Hill-Climbing algorithm to learn the structure of the graph. The generated graphical causal model is then presented to the operators via an adaptive and responsive user interface, where the users can give feedback to the system by reporting a correct or incorrect inference of the root cause, or by directly editing the generated model. Both the feedback and the changes to the model are used in the next iteration of the model generation.

Language-Based Assistance for Navigation

In the case of a vision-based navigation scenario, in which a human can request an agent to navigate and find a target object indoors, using just high-level language end-goals, the agent can query the human advisor for language sub-goals when it is lost and cannot make progress [132]. To implement such a methodology, it is required to learn a navigation and help policy to constrain the number of help requests that can be made to the advisor and learn at what points it is more effective to make the help requests for the overall progress in the task.

Cobot Motion Planning and Coordination Learning

A particular set of works focused on cobot motion planning and coordination learning tasks, where the human can be of great assistance to the machine by interacting with it and providing feedback on the expected collaborative robot behavior. Three works were retrieved that proposed human-in-the-loop assistance solutions in the form of demonstrations for cobot motion learning, and two others proposed more general frameworks to model and learn HRC processes through demonstrations.

Ref. [128] used learning from demonstrations (LfD) to perform robot motion synthesis with obstacle avoidance, using probabilistic task representations to estimate human occupancy in a long-term scale and a Hidden Semi-Markov Model (HSMM) to encode the demonstrated motions. An additional short-term prediction of human motion was added to account for sudden movements. The method was evaluated on reproducing air-drawing motions of letters near a moving human arm, with formal safety verification and replanning with a failsafe trajectory to avoid collision, resulting in no safety stops observed for all letters. The trajectories of the robot motion that were replanned to avoid collision still maintained the curvature and shape of the letters since the new trajectory was sampled from the probabilistic encoding of the movement. As the method does not need optimization or real-time collision checking, it computes in deterministic time; however, its validity and effectiveness are dependent on the human model used. A probabilistic extension of the method can be used to generate dynamic robot behavior, where the temporal dynamics are adapted by interactions with the user. The Adaptive Duration Hidden Semi-Markov Model [126] allows the robot to not only react to the user but to also behave proactively according to the encoded temporal coherence of the task. The authors of the method tracked the robot's positions and the human collaborator's hand positions during the demonstrations, with three different human motion velocities. The results of a collaborative handover and transportation task showed that the duration of the robot trajectory is strongly correlated with the human hand motion in the interaction part of the task. Proactive behavior was also observed, where the robot carried out the task when the user was unsure of how to perform it, in order to show its intention and how to proceed. Alternatively, dynamic time warping can be employed to process demonstrations with varying operation speeds and align their timelines. Ref. [134] used the method to teach a robotic arm assembly motion from human demonstrations. The proposed multi-dimensional dynamic time warping (SMMD-STW) method was able to segment and align multiple repetitions of teaching data with varying operation speeds, and a Mixture Gaussian regression (GMR) model was subsequently used to obtain an optimal reference trajectory and expected force field. The method was evaluated with an assembly task, showing lower force feedback compared to the teaching data due to the optimization and smoothing of the GMR model, and the admittance control applied by the controlling system. The method was evaluated with an assembly task, showing lower force feedback compared to the teaching data due to optimization and smoothing by the GMR model, and the admittance control applied by the controlling system.

Other models have been applied for learning HRC processes and coordination with human collaborators from demonstrations, namely, a Vector Autoregressive Partially Observable Markov Decision Process (VAR-POMDP) model [133], an extension of the POMDP method capable of modeling the dynamic system properties through the correlation be-

tween observations, and Deep Reinforcement Learning using a Deep Q-network model to approximate the optimal Q-function and learn the optimal policy [130].

Robot Control Learning

In the following four papers, robot control was learned from demonstrations and interventions, through the learning of expert policies and rewards.

In [131], the policy for an automated driving system was learned using an Interaction Imitation Learning method, a probabilistic variant of the DAGGER algorithm [137], to maximize sampling from a novice policy during learning (instead of the human expert), while constraining the probability of failure by a learned safety threshold. The Human-Gated DAGGER variant was proposed to give the human expert exclusive control of whether he/she or the novice should be in control at each moment. The novice policy was trained with an interactively augmented training dataset with labels collected from the expert during recovery control of the system. The risk metric was approximated by modeling the novice as an ensemble of neural networks and using the covariance matrix of the ensemble outputs to compute the policy confidence over the state space. The threshold to this risk metric at which control should be given back to the expert in unsafe situations was learned by computing the mean of the novice doubt at the time of human intervention (restricted to when the policy has already been trained with a lot of data and resembles the final fully trained policy). The method was evaluated in a simulated and real-work driving task. The results showed that the policies trained with HG-DAGGER method outperformed the ones trained with DAGGER and behavioral cloning. The performance, measured by road departure and collision rates, was better and learned faster and in a stabler way with HG-DAGGER. Moreover, the performance metrics were better when initializing the novice policy inside of the estimated safe and permissible set of states, based on the learned safety doubt threshold.

The goal of minimizing expert effort during human-in-the-loop learning was also addressed by [135] . The authors aimed at learning a shared control strategy for teleoperation in an online fashion from demonstrations and with reinforcement learning. The approach predicts the success probability in selecting one of the robot controllers and requesting human teleoperation if necessary. A multi-armed bandit (MAB) algorithm was employed to choose the best controller to use. MAB estimates the reward distribution for each arm from multiple trials, while minimizing the time cost imposed for requesting the human for demonstrations or for needing the human for failure recovery. The controller's policies were learned from demonstrations and continuously improved using online learning. The Deep Deterministic Policy Gradients algorithm was used for off-policy, model-free reinforcement learning of the robot policy. LfD was integrated into the DDPG framework with behavior cloning and a Q-filter. Demonstration of episodes by the controllers was added to a demonstration replay buffer. An additional behavior cloning loss term was then applied only to the samples of this buffer. The Q-filter ensured that the loss was not applied when the learned policy was significantly better than the demonstrated policy. The method was tested with two simulated tasks (block pushing and simulated navigation) and a real-world navigation task, using three controllers: a human teleoperator C_h, a pre-programmed baseline controller C_b, and a control policy learned online as described above C_l. The empirical results from the comparison with other approaches showed that the method led to a reduction in the total human cost. The results also showed that a better policy was learned using demonstrations from different sources, indicating a potential future application in learning from multiple dissimilar controllers.

Two works employed instead Inverse Reinforcement Learning (IRL). IRL can be more robust to changes in the tasks compared to Deep RL and learning from LfD, due to the difficulty of designing a suitable reward function. Moreover, it was also proposed to deal with the current problems of learning from expert demonstrations, namely, data sparsity that does not allow to efficiently learn complex sequential tasks, the cost of human interaction, and the reduced efficiency of learning from full demonstrations when the agent might just

be struggling to learn a specific part of the task. Ref. [136] used demonstrations and IRL to learn variable impedance control and reward functions for contact-rich manipulation tasks. The paper investigated the best action space for the reward function: whether rewards are defined for the force or impedance gain. Specifically, adversarial Inverse Reinforcement Learning was used to learn the impedance gain-based expert policy and reward function (the reward function was learned with the discriminator and the variable impedance policy was the generator), which was tested in a simulated and a real industrial robot experiment, achieving better transfer performance than the baselines. In [129], complex robot tasks were learned using a Human-in-the-loop Inverse Reinforcement Learning (HI-IRL) framework. The approach was more efficient than traditional IRL methods, and involved structured learning from failure experiences, from critical sub-goal information provided by an expert and partial demonstrations of sub-tasks that the agent struggled to learn. Experiments were performed in a grid world and car parking environment, showing that the method is more efficient, requiring fewer demonstrations than the baseline models to learn a task.

Using human feedback, demonstrations are commonly performed assuming they come from an expert and the information provided is optimal. The effectiveness of human-in-the-loop learning is, however, impacted by human cognitive bias, subjective preferences, and the quality of the human input which depends on the interaction/feedback modality used [23]. Humans can be considered non-stable, non-linear, and complex agents, affected by fatigue and other organizational and personal factors; therefore, the quality of human input should take this into account [5].

3.2.2. Human Assistance During Deployment

Only two research works were retrieved employing human-in-the-loop assistance during deployment (Table 11). In these works, the humans' capabilities are used by the AI algorithm during deployment to achieve better performance than the one that could be achieved without it. This type of collaborative intelligence can be likewise important to ensure the systems are functioning properly in production environments, according to the safety requirements, and respecting social and ethical guidelines.

Table 11. Description of the retrieved articles that used AI for CI tasks, where the human assists during the deployment phase of intelligent systems, ordered chronologically.

CI Interaction Type	Ref.	CI Task	AI Problem Domain	AI Method(s)	AI Paradigm	Application Domain
Human Assists Machine						
Human assistance during deployment	[138]	Hands free detection of emergency by the worker	Perception	DT classifier	Machine Learning	Manufacturing industry
	[139]	Gaze-assisted visual grounding for human–robot instruction	Communication	Faster R-CNN, Transformer-based text encoder, Text classifier	Machine Learning	Manufacturing industry
	[140]	Active real-time safety assessment	Perception	Online active learning	Machine Learning	Maritime industry

The acronyms used have the following expansions: Decision Tree (DT), Region-based Convolutional Neural Network (R-CNN).

Systems may also be designed to exploit human capabilities and assistance under normal operating conditions, beyond the learning phase. The following works explored three distinct solutions.

In [138], hands-free detection of emergencies in HRC scenarios was achieved by utilizing the workers' sensing abilities, where humans indirectly assist in emergency prediction by using data from a mobile Electroencephalogram (EEG) sensor as input. EEG can be used for the fast detection of brain activity changes that occur when an operator senses a potential emergency, and enters an alert or stressed state. A Decision Tree (DT) classifier was proposed to both classify the emergency state and offer a visualization of the inferred

classification rules. The feasibility of detecting emergencies using mobile EEG data was confirmed with a reliable window of 250 ms, but further experiments are needed to ensure the high level of robustness and speed that is required for safety compliance.

In the work of [139], a different type of implicit assistance was given during the deployment of a robotic agent for human–robot communication. The gaze of the human was used to assist the robotic agent with visual grounding (locating the most relevant object in an image based on a natural language query) in carrying out human instructions for pick and place tasks. The approach used a Faster R-CNN model [141] for object detection, a transformer-based text encoder model for text description embedding, and a text classification model for scoring candidate locations in regard to the text descriptions. Gaze-tracking was performed using the Microsoft Hololens2 head-mounted system, which provides a world coordinate of the point the eye is looking at. The gaze centroid was then used to select the detected object with a higher confidence score and closer to the gaze point. The gaze input improved the target localization accuracy from 26% to 65% in the test dataset.

Safety-critical applications in complex, non-stationary environments, such as deep-sea manned submersibles, require real-time safety assessment. To deal with concept-drift and class imbalances, ref. [140] proposed an online active learning approach that incrementally updates the safety assessment model. A novel use of the broad learning system under active learning and a new query strategy for reduced annotation cost of the users were developed.

The solution proposed by Yang et al. (2020) [44], already presented in Section 3.1.1, can be partially categorized as human assistance during the deployment of an ADAS system that learns iteratively in real-time the desired path through repeated cooperation with the driver, but after sufficient iterations, the system takes over to assist the human driver.

4. Discussion

This survey analyzed recent research work based on the type of CI interaction studied, the CI task performed, and the type of AI techniques used to address the task. It was clear that due to the extensiveness of the CI domain, only a sample of the available relevant literature was analyzed. Simultaneously, we have to consider that "collaborative intelligence" is still an emerging concept and that it has yet to be broadened to other domains. Next, we present the key takeaways from our review, insights into future research trends and the developed safety recommendations for practitioners in the CI field.

As a big portion of the reviewed methods were developed for the manufacturing context or safety-critical applications, such as aviation and the automotive industry, the majority of the proposed solutions consider a single component of a whole complex system. The evaluation and validation stage of the methods was varied, ranging from testing in simulated environments and tasks, testing with standard test sets, lab-based experimental testing using representative use-cases of the target applications, and assessment in multiple real-world scenarios and environments (performed by a minority of the works). Even still, the quality and representativeness of the training and testing data were rarely assessed, and when mentioned, it was to highlight the need for more diverse datasets. Several papers made use of high-fidelity digital twins to accelerate testing time and improve the safety of testing procedures, particularly for solutions that involve humans.

Establishing benchmarks for such complex and fast-developing AI technology is challenging but particularly for safety-critical and human–machine collaborative systems, at the minimum, the already existing industrial standards for system safety and methods specific for model safety, robustness, and dependability should be used at the different stages of model development. The recent review of Mohseni et al. (2022) [142] can be used to aid the selection of specific ML safety techniques, which can be applied at the design stage for inherently safe design specification, the development stage for increased performance and robustness, and/or the deployment stage for run-time error detection. In addition, the integration of different system modules should account for how errors propagate downstream

from the initial perception component to the final decision component, requiring the use of AI methods and metrics to quantify model uncertainty and transmit it through the system.

There is still a clear need for general guidelines, with comprehensive paradigm-agnostic metrics and specific procedures, that can be followed by CI researchers and engineers to ensure the reliability and dependability of the whole system before deployment in open-world settings. As a multitude of AI paradigms and models can be applied to solve one specific problem, and as it is typically difficult to determine what type of approach is more suitable to which problem, we recommend that these guidelines follow and build upon the CI task categorization here proposed that groups multiple similar tasks and high-level collaborative goals, in order to develop procedures that are independent of the model chosen and instead conditional on the type of data used and expected output:

- The proposed solutions for cobot motion planning and task scheduling are typically limited by the type of workspace/environment that the methods are applicable to, the number of workers accounted for by the solution, the types of tasks that they can be applied to, and by the uncertainties of human behavior. Dynamic and flexible approaches, such as the one by Zhu et al. (2019) [40], can better deal with variations in the expected operations but are usually harder to implement and test for compliance with safety requirements. Future research for robot motion planning and scheduling for collaborative human–robot interactions should be focused on dynamic/adaptive planning methods, online learning for the continuous improvement of the models, personalized assistance to the human collaborator, and more natural communication channels between human and robots. A particularly interesting and novel avenue to deal with the uncertainty of human behavior, and increase the comfort and safety of human–robot interactions is to estimate and match the human's expected behavior of the robot. The exploration of this problem will be key to develop systems that can resolve conflicts between humans and machines, reach a common ground, and adjust tasks and goals accordingly [9].
- The use of human demonstrations to aid the machine or AI learning process during training has a wide range of applications in robotics and automation, including for manipulation tasks in manufacturing domains, and for navigation tasks of mobile robots and vehicles [143]. Learning from demonstrations is a promising and versatile approach for robots to learn complex tasks by learning to imitate a human expert. There are, however, several limitations of the current techniques, such as the assumption that all demonstrations are optimal and the reliance on human data, which have pushed the research into investigating and developing solutions that (a) quantify the human factors in LfD and try to increase the quality of demonstrations [129]; (b) optimize the learning process while minimizing the human time cost of requesting demonstrations [135]; and (c) learn to generalize to novel scenarios, estimate the suitability to the new scenario, and determine when it needs human assistance/intervention [131,143]. As some of the reviewed works proposed, human assistance/intervention can be applied with less effort by training a model to perform the human's intervention job (can be either to reward or block agent actions), or by estimating, during real-time deployment, the confidence/doubt of the learned policy model relative to the execution risk, further improving safety [24]. Future applications should employ methods for human feedback quality assessment, human error detection, or outlier/out-of-distribution human sample detection, such as reducing data labeling ambiguity [144], improving demonstration quality scoring [145], or improving the human's response time in interventions that can lead to undesirable effects.
- The main limitation of the current human/object detection and tracking techniques is still robustness. As such, we suggest the use of (a) transfer learning, (b) ensemble classifiers, or (c) multiple sensor modalities to reduce occlusions and overlap effects, or (d) to take into consideration target variants/target changes of appearance as possible strategies to increase the robustness of detection systems. Future research should also

focus on developing novel methods to deal with unaccounted variability in the data after deployment, such as continual learning.

- The proposed solutions for human intention prediction in driving applications were limited by the fact that the developed improvements or sub-systems need to be integrated with an autonomous driving system/ADAS, and tested for robustness with real vehicles and environments. On the other hand, most of the motion recognition works proposed for human–robot collaboration applications were limited by the type of task or case study data used to train and test the models, as most focused only on single-person detection and upper limb movement data. Compared to the works that used image/video data, the papers that employed physiological data for motion intention recognition had comparatively small training datasets with a limited number of participants, even though they allowed to detect the intention up to 513 ms–2 s before the execution of the motion. One possible future work line for human motion intention prediction is the adaptation and personalization of the models to specific subjects when it becomes possible to have access to large amounts of big data that physiological sensors can provide. Fusing multiple modalities can also help to better distinguish between assistive or disruptive operator intentions, and intentional or incidental contact between humans and robots as in the work of Amin et al. (2020) [81], by combining vision and tactile perception data. With these limitations in mind, we recommend, as a future work line, going beyond the prediction of the driver's intentions for real-time assistance, and developing human-inspired safety metrics that can be retrieved from the trained models and serve as a prior for an autonomous driving policy.
- Ground truth labels for human state recognition can be obtained through subjective questionnaires or objective indicators. When these methods are not available during the online stage of the model implementation, behavioral indicators of the state of interest can be used instead (as in [96]) for continuous learning or model performance assessment. In addition, we recommend that systems that are already using wearable sensors to monitor human data for other purposes, such as using human posture and muscle activity for hand-over intention recognition [76], or human skeletal data for determining the optimal tool delivery position [48] could add a worker physical state recognition module for real-time robot behavior adaptation or to improve the ergonomic design of collaborative systems. Safety applications of human state recognition were addressed by the set of reviewed papers, demonstrating how these techniques can be used to prevent critical situations when the operator/user is not performing at the required level, and can cause harm to himself or others.

The challenges of building Hybrid Intelligent Systems include collaboration between humans and machines, adaptation to humans and the environment, systems that behave responsibly, and systems that can explain and share knowledge and strategies with each other [9]. Research has mainly focused on collaboration and adaptation, but more work is needed to develop explainable systems that take into account the ethical, legal, and societal responsibilities, as well as the consequences of these systems to better manage conflicts between intelligent systems and human experts.

Some of the reviewed CI solutions have shown signs of moving in this direction, such as the work of Oh et al. (2021) [55] and Li et al. (2022) [57], that proposed shared-control methods for the blending of the intelligent system's control commands and human commands, with the possibility to assign control authority to a human in high-risk scenarios or in cases of considerable disagreements between the two policies. This type of control solution, that is a middle-ground between direct and autonomous control, may be a suitable option for new teleoperation and autonomous driving systems in Industry 5.0, to comply with human agency and oversight requirements recently entered into force by the EU AI Act (Regulation (EU) 2024/1689) for high-risk AI systems. Still, other current machinery safety requirements, such as those of the Machinery Directive update (Regulation (EU) 2023/1230), require that any machinery, or related product with self-evolving behavior or logic that operates with varying levels of autonomy should communicate its planned

actions (such as what it is going to do and why) to operators in a comprehensible manner. This topic is rarely addressed in shared-control research, and methods from the explainable AI domain [146] and human–machine interaction domain [58] should be integrated and adopted. As an example, the presented work of Kottinger et al. (2022) [114] provided an explainability method for robot trajectory planning oversight, by generating human-understandable path visualizations.

Other CI solutions for conflict resolution may include, for when the intelligent system detects human decision-making bias or model disparities between the human and system, generating explanations that persuade the worker to take the most beneficial action or to repair their policy [105]. Alternatively, the estimate of the system's confidence/uncertainty on their outputs can be communicated to the human stakeholders to support informed decision-making [109,110].

An aspect of collaborative intelligence that was missed in the reviewed works is the fact that in many cyber–physical systems, the intelligent machine might have to interact with many humans in different contexts, and the humans might interact with different system instantiations, sharing the same knowledge base [2]. Computational techniques such as federated learning (also referred as collaborative learning) need to come into play to achieve this multi-human multi-agent collaboration scenario.

We provide a summary of recommended actions (Figure 7) taken from the analysis of the literature on the safe application of AI to collaborative intelligence problems.

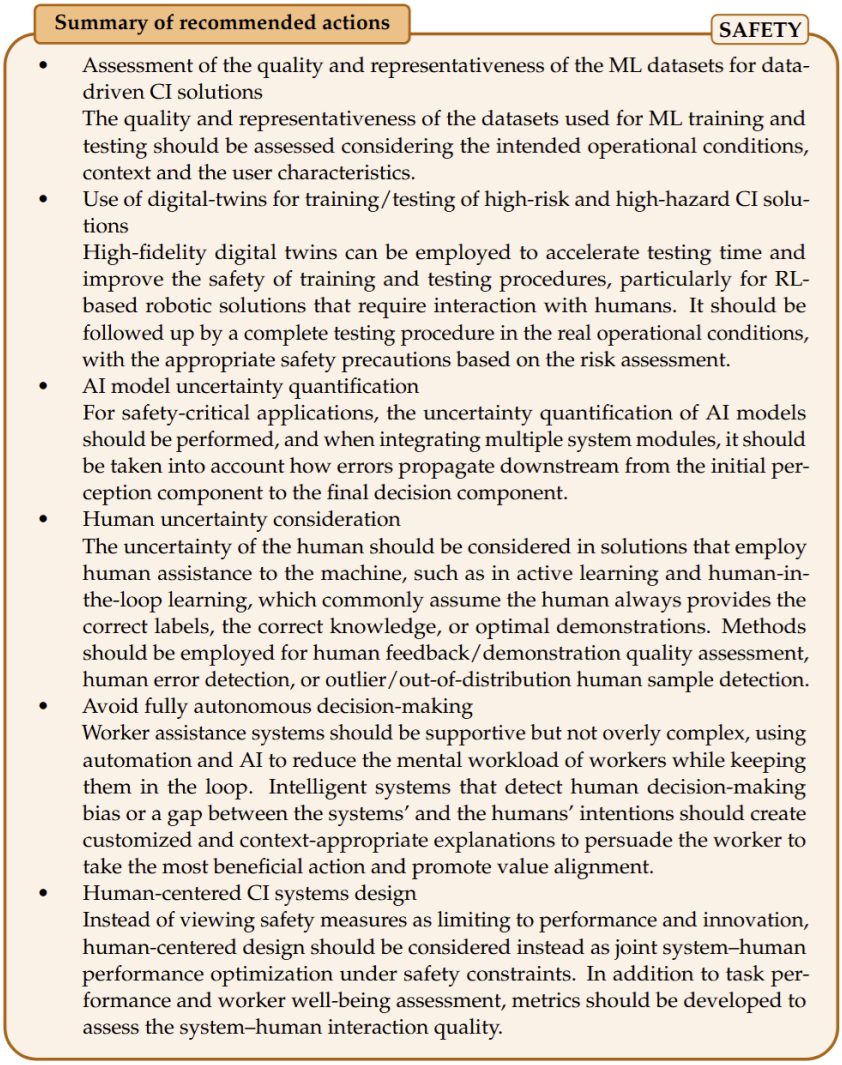

Figure 7. CI interaction types and task categories.

5. Conclusions

There are many opportunities in the collaborative intelligence field to take advantage of AI techniques that can help improve the interaction between humans, computers, and machines. In this paper, we reviewed a wide range of research work published in the last nine years proposing intelligent solutions for collaborative intelligence problems. The review focused on industrial applications that can benefit the most from human–machine teaming to achieve flexible and cost-effective automation solutions, and safety-critical industries that require human supervision as an additional safety layer for highly automated systems.

The retrieved research was organized into eleven common CI topics and further analyzed to gather insights into the type of AI problems that exist in CI applications and the type of AI algorithms used to solve the presented challenges. We then identified and discussed the key takeaways of the analysis process.

Lastly, we provided some recommendations for future research lines within the field of CI, with the aim of bridging the commonalities between the different topics that were addressed and achieving a more comprehensive and contextualized CI research landscape. Despite the advances in this field, it is still far from achieving true collaboration between humans and AI-based systems that is proactive and involves purposeful interactions, adaptation to context, and an understanding of each other's actions towards cooperative learning and problem-solving [9].

Author Contributions: Conceptualization, I.F.R. and M.C.L.; methodology, I.F.R. and M.C.L.; resources, E.D.; data curation, I.F.R.; writing—original draft preparation, I.F.R. and G.G.; writing—review and editing, I.F.R., M.C.L. and G.G.; visualization, I.F.R.; supervision, G.G., M.C.L. and E.D.; project administration, E.D.; funding acquisition, M.C.L., G.G. and E.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the CISC project (Marie Skłodowska-Curie grant agreement no. 955901). The work was partially supported by the MUSA-Multilayered Urban Sustainability Action project, funded by the European Union-NextGenerationEU, under the Mission 4 Component 2 Investment Line of the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D" (CUP G43C22001370007, Code ECS00000037); Program "piano sostegno alla ricerca" PSR and the PSR-GSA-Linea 6; Project ReGAInS (code 2023-NAZ-0207/DIP-ECC-DISCO23), funded by the Italian University and Research Ministry, within the Excellence Departments program 2023-2027 (law 232/2016); and-FAIR-Future Artificial Intelligence Research-Spoke 4-PE00000013-D53C22002380006, funded by the European Union-Next Generation EU within the project NRPP M4C2, Investment 1.,3 DD. 341, 15 March 2022.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hoc, J.M. From human-machine interaction to human-machine cooperation. *Ergonomics* **2000**, *43*, 833–843. [CrossRef] [PubMed]
2. Sendhoff, B.; Wersing, H. Cooperative Intelligence-A Humane Perspective. In Proceedings of the 2020 IEEE International Conference on Human-Machine Systems, ICHMS 2020, Rome, Italy, 7–9 September 2020. [CrossRef]
3. Kim, D.B.; Bajestani, M.S.; Lee, J.Y.; Shin, S.-J.; Noh, S.D. Introduction of Human-in-the-Loop in Smart Manufacturing (H-SM). *Int. J. Precis. Eng. Manuf.-Smart Tech.* **2024**, *2*, 209–214. [CrossRef]
4. Valtonen, L.; Mäkinen, S.J. Human-in-the-loop: Explainable or accurate artificial intelligence by exploiting human bias? In Proceedings of the 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference, Nancy, France, 19–23 June 2022; pp. 1–8. [CrossRef]
5. Ma, L.; Wang, C. Safety Issues in Human-Machine Collaboration and Possible Countermeasures. In Proceedings of the Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Anthropometry, Human Behavior, and Communication, Virtual Event, 26 June–1 July 2022; pp. 263–277.

6. High-Level Expert Group on AI (AI HLEG). In *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment | Shaping Europe's Digital Future*; European Commission: Luxembourg, 2020.
7. Bettoni, A.; Matteri, D.; Montini, E.; Gładysz, B.; Carpanzano, E. An AI adoption model for SMEs: A conceptual framework. *IFAC-PapersOnLine* **2021**, *54*, 702–708. [CrossRef]
8. Alix, C.; Lafond, D.; Mattioli, J.; Heer, J.D.; Chattington, M.; Robic, P.O. Empowering Adaptive Human Autonomy Collaboration with Artificial Intelligence. In Proceedings of the 2021 16th International System of Systems Engineering Conference, SoSE 2021, Vasteras, Sweden, 14–18 June 2021; pp. 126–131. [CrossRef]
9. Akata, Z.; Balliet, D.; De Rijke, M.; Dignum, F.; Dignum, V.; Eiben, G.; Fokkens, A.; Grossi, D.; Hindriks, K.; Hoos, H.; et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* **2020**, *53*, 18–28. [CrossRef]
10. *EN ISO 12100*; Safety of Machinery—General Principles for Design—Risk Assessment and Risk Reduction. International Organization for Standardization: Geneve, Switzerland, 2010.
11. *EN ISO 10218-2*; Robots and Robotic Devices—Safety Requirements for Industrial Robots Part 2: Robot Systems and Integration. International Organization for Standardization: Geneve, Switzerland, 2011.
12. *EN 50716*; Railway Applications—Requirements for Software Development. European Committee for Standardization (CEN): Brussels, Belgium, 2023.
13. *EN ISO 26262*; Road Vehicles—Functional Safety. International Organization for Standardization: Geneve, Switzerland, 2018.
14. *EN ISO 21448*; Road Vehicles—Safety of the Intended Functionality. International Organization for Standardization: Geneve, Switzerland, 2022.
15. *UL 4600*; Standard for Evaluation of Autonomous Products. American National Standards Institute (ANSI): Washington, DC, USA, 2023.
16. *EN ISO 9241-210*; Ergonomics of Human-System Interaction Part 210: Human-Centred Design for Interactive Systems. International Organization for Standardization: Geneve, Switzerland, 2019.
17. Hua, J.; Zeng, L.; Li, G.; Ju, Z. Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors* **2021**, *21*, 1278. [CrossRef]
18. Borboni, A.; Reddy, K.V.V.; Elamvazuthi, I.; AL-Quraishi, M.S.; Natarajan, E.; Azhar Ali, S.S. The Expanding Role of Artificial Intelligence in Collaborative Robots for Industrial Applications: A Systematic Review of Recent Works. *Machines* **2023**, *11*, 111. [CrossRef]
19. Liu, H.; Wang, L. Gesture recognition for human-robot collaboration: A review. *Int. J. Ind. Ergon.* **2018**, *68*, 355–367. [CrossRef]
20. Zhang, R.; Saran, A.; Liu, B.; Zhu, Y.; Guo, S.; Niekum, S.; Ballard, D.; Hayhoe, M. Human gaze assisted artificial intelligence: A review. *Ijcai Int. Jt. Conf. Artif. Intell.* **2020**, *2021*, 4951–4958. [CrossRef]
21. Šumak, B.; Brdnik, S.; Pušnik, M. Sensors and artificial intelligence methods and algorithms for human–computer intelligent interaction: A systematic mapping study. *Sensors* **2022**, *22*, 20. [CrossRef]
22. Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; He, L. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* **2022**, *135*, 364–381. [CrossRef]
23. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artif. Intell. Rev.* **2023**, *56*, 3005–3054. [CrossRef]
24. Rajendran, P.T.; Espinoza, H.; Delaborde, A.; Mraidha, C. Human-in-the-Loop Learning Methods Toward Safe DL-Based Autonomous Systems: A Review. In *Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops. SAFECOMP 2021*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; pp. 251–264. [CrossRef]
25. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The Limitations of Deep Learning in Adversarial Settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbrucken, Germany, 21–24 March 2016; pp. 372–387. [CrossRef]
26. Gianini, G.; Guzzo, L. Artificial Intelligence and Machine Learning: An Introduction—Astrofisica Alla Statale. Available online: https://astro.fisica.unimi.it/people/luigi-guzzo/ml-intro/ (accessed on 5 July 2022).
27. HELGA. *In A Definition of AI: Main Capabilities and Scientific Disciplines*; European Commission: Luxembourg, 2019.
28. Corea, F. *AI Knowledge Map: How to Classify AI Technologies*; Springer: Berlin/Heidelberg, Germany, 2019. [CrossRef]
29. Calegari, R.; Ciatto, G.; Denti, E.; Omicini, A. Logic-based technologies for intelligent systems: State of the art and perspectives. *Information* **2020**, *11*, 167. [CrossRef]
30. Lakemeyer, G.; Nebel, B. Foundations of knowledge representation and reasoning: A guide to this volume. In *Foundations of Knowledge Representation and Reasoning*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 1994; Volume 810, pp. 1–12. [CrossRef]
31. Fischer, A.; Igel, C. An introduction to restricted Boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012;* Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7441, pp. 14–36. [CrossRef]
32. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 29, pp. 1–73.
33. Kelleher, J.D. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2019.

34. Fernandes Ramos, I.F.; Gianini, G.; Damiani, E. Neuro-Symbolic AI for Sensor-based Human Performance Prediction: Challenges and Solutions. In Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022), Dublin, Ireland, 28 August–1 September 2022. [CrossRef]
35. Venter, G. Review of Optimization Techniques. *Encycl. Aerosp. Eng.* **2010**, 1–12. [CrossRef]
36. Koditschek, D.E. What Is Robotics? Why Do We Need It and How Can We Get It? *Annu. Rev. Control. Robot. Auton. Syst.* **2021**, *4*, 1–33. [CrossRef]
37. Pellegrinelli, S.; Orlandini, A.; Pedrocchi, N.; Umbrico, A.; Tolio, T. Motion planning and scheduling for human and industrial-robot collaboration. *CIRP Ann.-Manuf. Technol.* **2017**, *66*, 1–4. [CrossRef]
38. Guo, M.; Andersson, S.; Dimarogonas, D.V. Human-in-the-Loop Mixed-Initiative Control under Temporal Tasks. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 6395–6400. [CrossRef]
39. Cesta, A.; Orlandini, A.; Umbrico, A. Fostering Robust Human-Robot Collaboration through AI Task Planning. *Proc. Procedia Cirp* **2018**, *72*, 1045–1050. [CrossRef]
40. Zhu, L.; Chi, Z.; Zhou, F.; Zhuang, C. Dynamic motion planning algorithm in human-robot collision avoidance. In *Intelligent Robotics and Applications, Proceedings of the 12th International Conference, ICIRA 2019, Shenyang, China, 8–11 August 2019*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11745, pp. 655–666. [CrossRef]
41. Kulkarni, A.; Vadlamudi, S.G.; Zha, Y.; Zhang, Y.; Chakraborti, T.; Kambhampati, S. Explicable planning as minimizing distance from expected behavior. In Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, London, UK, 29 May–2 June 2023; Volume 4, pp. 2075–2077.
42. Koert, D.; Pajarinen, J.; Schotschneider, A.; Trick, S.; Rothkopf, C.; Peters, J. Learning Intention Aware Online Adaptation of Movement Primitives. *IEEE Robot. Autom. Lett.* **2019**, *4*, 3719–3726. [CrossRef]
43. El-Shamouty, M.; Wu, X.; Yang, S.; Albus, M.; Huber, M.F. Towards Safe Human-Robot Collaboration Using Deep Reinforcement Learning. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 4899–4905. [CrossRef]
44. Yang, L.; Li, Y.; Huang, D.; Xia, J. Iterative learning of an unknown road path through cooperative driving of vehicles. *IET Intell. Transp. Syst.* **2020**, *14*, 423–431. [CrossRef]
45. Yang, X.; Cheng, J.; Michael, N. An Intention Guided Hierarchical Framework for Trajectory-based Teleoperation of Mobile Robots. In Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 13589–13595. [CrossRef]
46. Ionescu, T.B. Adaptive simplex architecture for safe, real-time robot path planning. *Sensors* **2021**, *21*, 2589. [CrossRef]
47. Lv, Q.; Zhang, R.; Sun, X.; Lu, Y.; Bao, J. A digital twin-driven human-robot collaborative assembly approach in the wake of COVID-19. *J. Manuf. Syst.* **2021**, *60*, 837–851. [CrossRef]
48. Khawaja, F.I.; Kanazawa, A.; Kinugawa, J.; Kosuge, K. A human-following motion planning and control scheme for collaborative robots based on human motion prediction. *Sensors* **2021**, *21*, 8229. [CrossRef] [PubMed]
49. Eckhoff, M.; Knobbe, D.; Zwirnmann, H.; Swikir, A.; Haddadin, S. Towards Connecting Control to Perception: High-Performance Whole-Body Collision Avoidance Using Control-Compatible Obstacles. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 2354–2361. [CrossRef]
50. Tausch, A.; Kluge, A.; Adolph, L. Psychological Effects of the Allocation Process in Human–Robot Interaction—A Model for Research on ad hoc Task Allocation. *Front. Psychol.* **2020**, *11*, 564672. [CrossRef] [PubMed]
51. Onnasch, L.; Wickens, C.D.; Li, H.; Manzey, D. Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Hum. Factors* **2014**, *56*, 476–488. [CrossRef] [PubMed]
52. Zabihifar, S.; Yuschenko, A. Hybrid force/position control of a collaborative parallel robot using adaptive neural network. In *Interactive Collaborative Robotics, Proceedings of the Third International Conference, ICR 2018, Leipzig, Germany, 18–22 September 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11097, pp. 280–290. [CrossRef]
53. Roveda, L.; Haghshenas, S.; Caimmi, M.; Pedrocchi, N.; Tosatti, L.M. Assisting operators in heavy industrial tasks: On the design of an optimized cooperative impedance fuzzy-controller with embedded safety rules. *Front. Robot. AI* **2019**, *6*, 463524. [CrossRef] [PubMed]
54. Zhao, X.; Fan, T.; Li, Y.; Zheng, Y.; Pan, J. An Efficient and Responsive Robot Motion Controller for Safe Human-Robot Collaboration. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6068–6075. [CrossRef]
55. Oh, Y.; Toussaint, M.; Mainprice, J. Learning to Arbitrate Human and Robot Control using Disagreement between Sub-Policies. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 27 September–1 October 2021; pp. 5305–5311. [CrossRef]
56. Costanzo, M.; De Maria, G.; Lettera, G.; Natale, C. A Multimodal Approach to Human Safety in Collaborative Robotic Workcells. *IEEE Trans. Autom. Sci. Eng.* **2021**, *19*, 1202–1216. [CrossRef]
57. Li, W.; Li, Q.; Li, S.E.; Li, R.; Ren, Y.; Wang, W. Indirect Shared Control Through Non-Zero Sum Differential Game for Cooperative Automated Driving. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15980–15992. [CrossRef]
58. Losey, D.P.; McDonald, C.G.; Battaglia, E.; O'Malley, M.K. A Review of Intent Detection, Arbitration, and Communication Aspects of Shared Control for Physical Human–Robot Interaction. *Appl. Mech. Rev.* **2018**, *70*, 010804. [CrossRef]

59. Maccarini, M.; Pura, F.; Piga, D.; Roveda, L.; Mantovani, L.; Braghin, F. Preference-Based Optimization of a Human-Robot Collaborative Controller. *IFAC-PapersOnLine* **2022**, *55*, 7–12. [CrossRef]
60. Fang, J.; Yuan, Y. Human-in-the-loop optimization of wearable robots to reduce the human metabolic energy cost in physical movements. *Robot. Auton. Syst.* **2020**, *127*, 103495. [CrossRef]
61. Perez, J.; Flores, J.L.; Blum, C.; Cerquides, J.; Abuin, A. Optimization Techniques and Formal Verification for the Software Design of Boolean Algebra Based Safety-Critical Systems. *IEEE Trans. Ind. Inform.* **2022**, *18*, 620–630. [CrossRef]
62. Giallanza, A.; La Scalia, G.; Micale, R.; La Fata, C.M. Occupational health and safety issues in human-robot collaboration: State of the art and open challenges. *Saf. Sci.* **2024**, *169*, 106313. [CrossRef]
63. Gualtieri, L.; Fraboni, F.; De Marchi, M.; Rauch, E. Development and evaluation of design guidelines for cognitive ergonomics in human-robot collaborative assembly systems. *Appl. Ergon.* **2022**, *104*, 103807. [CrossRef] [PubMed]
64. Kokotinis, G.; Michalos, G.; Arkouli, Z.; Makris, S. On the quantification of human-robot collaboration quality. *Int. J. Comput. Integr. Manuf.* **2023**, *36*, 1431–1448. [CrossRef]
65. Lalos, C.; Voulodimos, A.; Doulamis, A.; Varvarigou, T. Efficient tracking using a robust motion estimation technique. *Multimed. Tools Appl.* **2014**, *69*, 277–292. [CrossRef]
66. Maderna, R.; Lanfredini, P.; Zanchettin, A.M.; Rocco, P. Real-time monitoring of human task advancement. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Macau, China, 3–8 November 2019; pp. 433–440. [CrossRef]
67. Silva, I.R.; Barbosa, G.B.; Ledebour, C.C.; Filho, A.T.; Kelner, J.; Sadok, D.; Lins, S.; Souza, R. Assessing Deep Learning Models for Human-Robot Collaboration Collision Detection in Industrial Environments. In *Intelligent Systems, Proceedings of the 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, 20–23 October 2020*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12319, pp. 240–255. [CrossRef]
68. Fraga-Lamas, P.; Barros, D.; Lopes, S.I.; Fernández-Caramés, T.M. Mist and Edge Computing Cyber-Physical Human-Centered Systems for Industry 5.0: A Cost-Effective IoT Thermal Imaging Safety System. *Sensors* **2022**, *22*, 8500. [CrossRef]
69. Driggs-Campbell, K.; Bajcsy, R. Identifying Modes of Intent from Driver Behaviors in Dynamic Environments. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, Gran Canaria, Spain, 15–18 September 2015; pp. 739–744. [CrossRef]
70. Coupeté, E.; Moutarde, F.; Manitsaris, S. Gesture Recognition Using a Depth Camera for Human Robot Collaboration on Assembly Line. *Procedia Manuf.* **2015**, *3*, 518–525. [CrossRef]
71. Liu, C.; Hamrick, J.B.; Fisac, J.F.; Dragan, A.D.; Hedrick, J.K.; Sastry, S.S.; Griffiths, T.L. Goal inference improves objective and perceived performance in human-robot collaboration. In Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, New York, NY, USA, 19–23 July 23 2004; pp. 940–948.
72. Wang, P.; Liu, H.; Wang, L.; Gao, R.X. Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. *Cirp-Ann.-Manuf. Technol.* **2018**, *67*, 17–20. [CrossRef]
73. Zanchettin, A.M.; Casalino, A.; Piroddi, L.; Rocco, P. Prediction of Human Activity Patterns for Human-Robot Collaborative Assembly Tasks. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3934–3942. [CrossRef]
74. Wu, M.; Louw, T.; Lahijanian, M.; Ruan, W.; Huang, X.; Merat, N.; Kwiatkowska, M. Gaze-based Intention Anticipation over Driving Manoeuvres in Semi-Autonomous Vehicles. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Macau, China, 3–8 November 2019; pp. 6210–6216. [CrossRef]
75. Neto, P.; Simão, M.; Mendes, N.; Safeea, M. Gesture-based human-robot interaction for human assistance in manufacturing. *Int. J. Adv. Manuf. Technol.* **2019**, *101*, 119–135. [CrossRef]
76. Wang, W.; Li, R.; Diekel, Z.M.; Chen, Y.; Zhang, Z.; Jia, Y. Controlling object hand-over in human-robot collaboration via natural wearable sensing. *IEEE Trans. Hum.-Mach. Syst.* **2019**, *49*, 59–71. [CrossRef]
77. Luo, R.C.; Mai, L. Human Intention Inference and On-Line Human Hand Motion Prediction for Human-Robot Collaboration. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Macau, China, 4–8 November 2019; pp. 5958–5964. [CrossRef]
78. Casalino, A.; Brameri, A.; Zanchettin, A.M.; Rocco, P. Adaptive swept volumes generation for human-robot coexistence using Gaussian Processes. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Macau, China, 3–8 November 2019; pp. 3823–3829. [CrossRef]
79. Shih, J.C. Predicting stochastic human forward reachable sets based on learned human behavior. In Proceedings of the American Control Conference, Philadelphia, PA, USA, 10–12 July 2019; pp. 5247–5253. [CrossRef]
80. van Rooijen, S.J.; Ellerbroek, J.; Borst, C.; van Kampen, E. Toward individual-sensitive automation for air traffic control using convolutional neural networks. *Proc. J. Air Transp.* **2020**, *28*, 105–113. [CrossRef]
81. Amin, F.M.; Rezayati, M.; van de Venn, H.W.; Karimpour, H. A mixed-perception approach for safe human–robot collaboration in industrial automation. *Sensors* **2020**, *20*, 6347. [CrossRef] [PubMed]
82. Pham, D.T.; Alam, S.; Duong, V. An Air Traffic Controller Action Extraction-Prediction Model Using Machine Learning Approach. *Complexity* **2020**, *2020*, 1659103. [CrossRef]
83. Anvaripour, M.; Khoshnam, M.; Menon, C.; Saif, M. FMG- and RNN-Based Estimation of Motor Intention of Upper-Limb Motion in Human-Robot Collaboration. *Front. Robot. AI* **2020**, *7*, 573096. [CrossRef]
84. Buerkle, A.; Eaton, W.; Lohse, N.; Bamber, T.; Wolfson, P.F. EEG based arm movement intention recognition towards enhanced safety in symbiotic Human-Robot Collaboration. *Robot.-Comput.-Integr. Manuf.* **2021**, *70*, 102137. [CrossRef]

85. Elmalaki, S. FaiR-IoT: Fairness-Aware Human-in-the-Loop Reinforcement Learning for Harnessing Human Variability in Personalized IoT. In Proceedings of the Proceedings of the International Conference on Internet-of-Things Design and Implementation, New York, NY, USA, 18–21 May 2021; pp. 119–132. [CrossRef]
86. Dong, M.; Peng, J.; Ding, S.; Wang, Z. Transfer Learning-Based Intention Recognition of Human Upper Limb in Human—Robot Collaboration. In *Intelligent Robotics and Applications, Proceedings of the 14th International Conference, ICIRA 2021, Yantai, China, 22–25 October 2021*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; Volume 13014, pp. 586–595. [CrossRef]
87. Li, S.; Fan, J.; Zheng, P.; Wang, L. Transfer Learning-enabled Action Recognition for Human-robot Collaborative Assembly. *Proc. Procedia Cirp* **2021**, *104*, 1795–1800. [CrossRef]
88. Gomez Cubero, C.; Rehm, M. Intention Recognition in Human Robot Interaction Based on Eye Tracking. In *Human-Computer Interaction—INTERACT 2021, Proceedings of the 18th IFIP TC 13 International Conference, Bari, Italy, 30 August–3 September 2021*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; Volume 12934, pp. 428–437. [CrossRef]
89. Zhang, Z.; Han, S.; Yi, H.; Duan, F.; Kang, F.; Sun, Z.; Sole-Casals, J.; Caiafa, C.F. A Brain-Controlled Vehicle System Based on Steady State Visual Evoked Potentials. *Cogn. Comput.* **2023**, *15*, 159–175. [CrossRef]
90. Orsag, L.; Stipancic, T.; Koren, L. Towards a Safe Human–Robot Collaboration Using Information on Human Worker Activity. *Sensors* **2023**, *23*, 1283. [CrossRef]
91. Dimova-Edeleva, V.; Rivera, O.S.; Laha, R.; Figueredo, L.F.C.; Zavaglia, M.; Haddadin, S. Error-related Potentials in a Virtual Pick-and-Place Experiment: Toward Real-world Shared-control. In Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 24–27 July 2023; pp. 1–7. [CrossRef]
92. Popov, D.; Pashkevich, A.; Klimchik, A. Adaptive technique for physical human-robot interaction handling using proprioceptive sensors. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107141. [CrossRef]
93. McAllister, R.; Gal, Y.; Kendall, A.; Van Der Wilk, M.; Shah, A.; Cipolla, R.; Weller, A. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. *IJCAI Int. Jt. Conf. Artif. Intell.* **2017**, 4745–4753. [CrossRef]
94. Li, N.; Misu, T.; Miranda, A. Driver behavior event detection for manual annotation by clustering of the driver physiological signals. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, Rio de Janeiro, Brazil, 1-4 November 2016; pp. 2583–2588. [CrossRef]
95. Ma, Y.; Zhang, S.; Qi, D.; Luo, Z.; Li, R.; Potter, T.; Zhang, Y. Driving drowsiness detection with EEG using a modified hierarchical extreme learning machine algorithm with particle swarm optimization: A pilot study. *Electronics* **2020**, *9*, 775. [CrossRef]
96. Huang, H.; Cappel, H.F.; Thomas, G.C.; He, B.; Sentis, L. Adaptive Compliance Shaping with Human Impedance Estimation. In Proceedings of the American Control Conference, Denver, CO, USA, 1–3 July 2020; pp. 5131–5138. [CrossRef]
97. Zhang, X.; Mahadevan, S.; Lau, N.; Weinger, M.B. Multi-source information fusion to assess control room operator performance. *Reliab. Eng. Syst. Saf.* **2020**, *194*, 106287. [CrossRef]
98. Diamantopoulos, H.; Wang, W. Accommodating and assisting human partners in human-robot collaborative tasks through emotion understanding. In Proceedings of the 2021 12th International Conference on Mechanical and Aerospace Engineering, ICMAE 2021, Athens, Greece, 16–19 July 2021; IEEE: Piscataway, NJ, USA; pp. 523–528. [CrossRef]
99. John, A.R.; Singh, A.K.; Do, T.T.N.; Eidels, A.; Nalivaiko, E.; Gavgani, A.M.; Brown, S.; Bennett, M.; Lal, S.; Simpson, A.M.; et al. Unraveling the Physiological Correlates of Mental Workload Variations in Tracking and Collision Prediction Tasks. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 770–781. [CrossRef] [PubMed]
100. Alkinani, M.H.; Khan, W.Z.; Arshad, Q.; Raza, M. HSDDD: A Hybrid Scheme for the Detection of Distracted Driving through Fusion of Deep Learning and Handcrafted Features. *Sensors* **2022**, *22*, 1864. [CrossRef] [PubMed]
101. Sukhavasi, S.B.; Sukhavasi, S.B.; Elleithy, K.; El-Sayed, A.; Elleithy, A. A Hybrid Model for Driver Emotion Detection Using Feature Fusion Approach. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3085. [CrossRef]
102. Shayesteh, S.; Ojha, A.; Liu, Y.; Jebelli, H. Human-robot teaming in construction: Evaluative safety training through the integration of immersive technologies and wearable physiological sensing. *Saf. Sci.* **2023**, *159* . [CrossRef]
103. Görür, O.C.; Rosman, B.; Sivrikaya, F.; Albayrak, S. FABRIC: A Framework for the Design and Evaluation of Collaborative Robots with Extended Human Adaptation. *J. Hum.-Robot Interact.* **2023**, *12*, 1–54. [CrossRef]
104. Zanoli, C.; Villani, V.; Picone, M. The Road to Industry 5.0: The Challenges of Human Fatigue Modeling. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, HI, USA, 1–4 October 2023; pp. 4797–4804. [CrossRef]
105. Tabrez, A.; Agrawal, S.; Hayes, B. Explanation-Based Reward Coaching to Improve Human Performance via Reinforcement Learning. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Daegu, Republic of Korea, 1–14 March 2019; pp. 249–257. [CrossRef]
106. Blum, S.; Klaproth, O.; Russwinkel, N. Cognitive Modeling of Anticipation: Unsupervised Learning and Symbolic Modeling of Pilots' Mental Representations. *Top. Cogn. Sci.* **2022**, *14*, 718–738. [CrossRef]
107. Zhang, X.; Zhu, Y.; Lin, H. Performance guaranteed human-robot collaboration through correct-by-design. In Proceedings of the 2016 American Control Conference (ACC), Boston, MA, USA, 6–8 July 2016; pp. 6183–6188. [CrossRef]
108. Rajnathsing, H.; Li, C. A neural network based monitoring system for safety in shared work-space human-robot collaboration. *Ind. Robot.* **2018**, *45*, 481–491. [CrossRef]

109. Fisac, J.; Bajcsy, A.; Herbert, S.; Fridovich-Keil, D.; Wang, S.; Tomlin, C.; Dragan, A. Probabilistically Safe Robot Planning with Confidence-Based Human Predictions. In Proceedings of the Robotics: Science and Systems, Pittsburgh, PA, USA, 26–30 June 2018. [CrossRef]
110. Fridovich-Keil, D.; Bajcsy, A.; Fisac, J.F.; Herbert, S.L.; Wang, S.; Dragan, A.D.; Tomlin, C.J. Confidence-aware motion prediction for real-time collision avoidance1. *Int. J. Robot. Res.* **2020**, *39*, 250–265. [CrossRef]
111. Wu, B.; Zhang, X.; Lin, H. Supervisor synthesis of POMDP via automata learning. *Automatica* **2021**, *129*, 109654. [CrossRef]
112. Scalera, L.; Giusti, A.; Vidoni, R.; Gasparetto, A. Enhancing fluency and productivity in human-robot collaboration through online scaling of dynamic safety zones. *Int. J. Adv. Manuf. Technol.* **2022**, *121*, 6783–6798. [CrossRef]
113. Haindl, P.; Buchgeher, G.; Khan, M.; Moser, B. Towards a Reference Software Architecture for Human-AI Teaming in Smart Manufacturing. In Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results, New York, NY, USA, 22–24 May 2022; pp. 96–100. [CrossRef]
114. Kottinger, J.; Almagor, S.; Lahijanian, M. Conflict-Based Search for Explainable Multi-Agent Path Finding. In Proceedings of the International Conference on Automated Planning and Scheduling, ICAPS, Virtual, 13–24 June 2022; Volume 32, pp. 692–700. [CrossRef]
115. Lloyd-Roberts, B.; James, P.; Edwards, M.; Robinson, S.; Werner, T. Improving Railway Safety: Human-in-the-loop Invariant Finding. In Proceedings of the Conference on Human Factors in Computing Systems-Proceedings, Hamburg, Germany, 23–28 April 2023. [CrossRef]
116. Islam, S.O.B.; Lughmani, W.A. A Connective Framework for Safe Human–Robot Collaboration in Cyber-Physical Production Systems. *Arab. J. Sci. Eng.* **2023**, *48*, 11621–11644. [CrossRef]
117. Wang, K.J.; Lin, C.J.; Tadesse, A.A.; Woldegiorgis, B.H. Modeling of human-robot collaboration for flexible assembly-a hidden semi-Markov-based simulation approach. *Int. J. Adv. Manuf. Technol.* **2023**, *126*, 5371–5389. [CrossRef]
118. Angluin, D. Learning regular sets from queries and counterexamples. *Inf. Comput.* **1987**, *75*, 87–106. [CrossRef]
119. Zhang, R.; Furusho, M. Constructing a decision-support system for safe ship-navigation using a Bayesian network. In *Digital Human Modeling: Applications in Health, Safety, Ergonomics and Risk Management, Proceedings of the7th International Conference, DHM 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, 17–22 July 2016*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9745, pp. 616–628. [CrossRef]
120. Elmalaki, S.; Tsai, H.R.; Srivastava, M. Sentio: Driver-in-the-loop forward collision warning using multisample reinforcement learning. In Proceedings of the SenSys 2018-Proceedings of the 16th Conference on Embedded Networked Sensor Systems, Shenzhen China, 4–7 November 2018; pp. 28–40. [CrossRef]
121. Singh, H.V.; Mahmoud, Q.H. Human-in-the-Loop Error Precursor Detection using Language Translation Modeling of HMI States. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Toronto, ON Canada, 11–14 October 2020; pp. 2237–2242. [CrossRef]
122. Wang, Z.; Liao, X.; Wang, C.; Oswald, D.; Wu, G.; Boriboonsomsin, K.; Barth, M.J.; Han, K.; Kim, B.; Tiwari, P. Driver Behavior Modeling Using Game Engine and Real Vehicle: A Learning-Based Approach. *IEEE Trans. Intell. Veh.* **2020**, *5*, 738–749. [CrossRef]
123. Goldberg, K. Robots and the return to collaborative intelligence. *Nat. Mach. Intell.* **2019**, *1*, 2–4. [CrossRef]
124. Saunders, W.; Sastry, G.; Stuhlmüller, A.; Evans, O. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, Stockholm, Sweden, 10–15 July 2018; pp. 2067–2069.
125. Chen, J.; Wu, T.; Shi, M.; Jiang, W. PORF-DDPG: Learning personalized autonomous driving behavior with progressively optimized reward function. *Sensors* **2020**, *20*, 5626. [CrossRef]
126. Rozo, L.; Silvério, J.; Calinon, S.; Caldwell, D.G. Learning controllers for reactive and proactive behaviors in human-robot collaboration. *Front. Robot. AI* **2016**, *3*, 30. [CrossRef]
127. Büttner, S.; Wunderlich, P.; Heinz, M.; Niggemann, O.; Röcker, C. Managing Complexity: Towards Intelligent Error-Handling Assistance Trough Interactive Alarm Flood Reduction. In *Machine Learning and Knowledge Extraction, Proceedings of the First IFIP TC 5, WG 8.4, 8.9, 12.9 International Cross-Domain Conference, CD-MAKE 2017, Reggio, Italy, 29 August–1 September 2017*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10410, pp. 69–82. [CrossRef]
128. Zeestraten, M.J.; Pereira, A.; Althoff, M.; Calinon, S. Online motion synthesis with minimal intervention control and formal safety guarantees. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016-Conference Proceedings, Banff, AB, Canada, 5–8 October 2017; pp. 2116–2121. [CrossRef]
129. Pan, X.; Shen, Y. Human-Interactive Subgoal Supervision for Efficient Inverse Reinforcement Learning. In Proceedings of the AAMAS '18: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, Stockholm, Sweden, 10–15 July 2018; pp. 1380–1387. [CrossRef]
130. Lombardi, M.; Liuzza, D.; Bemardo, M.D. Deep learning control of artificial avatars in group coordination tasks. In Proceedings of the-IEEE International Conference on Systems, Man and Cybernetics, Bari, Italy, 6–9 October 2019; pp. 714–719. [CrossRef]
131. Kelly, M.; Sidrane, C.; Driggs-Campbell, K.; Kochenderfer, M.J. HG-DAgger: Interactive imitation learning with human experts. In Proceedings of the IEEE International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 8077–8083. [CrossRef]

132. Nguyen, K.; Dey, D.; Brockett, C.; Dolan, B. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 20 June 2019; pp. 12519–12529. [CrossRef]
133. Zheng, W.; Lin, H. Vector Autoregressive POMDP Model Learning and Planning for Human–Robot Collaboration. *IEEE Control Syst. Lett.* **2019**, *3*, 775–780. [CrossRef]
134. Lin, H.I. Design of an intelligent robotic precise assembly system for rapid teaching and admittance control. *Robot. Comput.-Integr. Manuf.* **2020**, *64*, 101946. [CrossRef]
135. Rigter, M.; Lacerda, B.; Hawes, N. A Framework for Learning from Demonstration with Minimal Human Effort. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2023–2030. [CrossRef]
136. Zhang, X.; Sun, L.; Kuang, Z.; Tomizuka, M. Learning Variable Impedance Control via Inverse Reinforcement Learning for Force-Related Tasks. *IEEE Robot. Autom. Lett.* **2021**, *6*, 2225–2232. [CrossRef]
137. Ross, S.; Gordon, G.J.; Bagnell, J.A. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; Volume 15. [CrossRef]
138. Buerkle, A.; Bamber, T.; Lohse, N.; Ferreira, P. Feasibility of Detecting Potential Emergencies in Symbiotic Human-Robot Collaboration with a mobile EEG. *Robot. Comput.-Integr. Manuf.* **2021**, *72*, 102179. [CrossRef]
139. Johari, K.; Tong, C.T.Z.; Subbaraju, V.; Kim, J.J.; Tan, U.X. Gaze Assisted Visual Grounding. In *Social Robotics, Proceedings of the 13th International Conference, ICSR 2021, Singapore, 10–13 November 2021*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; Volume 13086, pp. 191–202. [CrossRef]
140. Liu, Z.; Zhang, Y.; Ding, Z.; He, X. An Online Active Broad Learning Approach for Real-Time Safety Assessment of Dynamic Systems in Nonstationary Environments. *IEEE Trans. Neural Networks Learn. Syst.* **2023**, *34*, 6714–6724. [CrossRef]
141. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
142. Mohseni, S.; Wang, H.; Xiao, C.; Yu, Z.; Wang, Z.; Yadawa, J. Taxonomy of Machine Learning Safety: A Survey and Primer. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
143. Ravichandar, H.; Polydoros, A.S.; Chernova, S.; Billard, A. Recent Advances in Robot Learning from Demonstration. *Annu. Rev. Control. Robot. Auton. Syst.* **2020**, *3*, 297–330. [CrossRef]
144. Deng, Z.; Li, T.; Deng, D.; Liu, K.; Luo, Z.; Zhang, P. Feature Selection for Handling Label Ambiguity Using Weighted Label-Fuzzy Relevancy and Redundancy. *IEEE Trans. Fuzzy Syst.* **2024**, *32*, 4436–4447. [CrossRef]
145. Sakr, M.; Li, Z.J.; Van der Loos, H.F.M.; Kulić, D.; Croft, E.A. Quantifying Demonstration Quality for Robot Learning and Generalization. *IEEE Robot. Autom. Lett.* **2022**, *7*, 9659–9666. [CrossRef]
146. Hagras, H. Toward Human-Understandable, Explainable AI. *Computer* **2018**, *51*, 28–36. [CrossRef]