IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 February 2023, accepted 20 April 2023, date of publication 28 April 2023, date of current version 9 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3271602

SURVEY

Benchmarking Teamwork of Humans and Cobots–An Overview of Metrics, Strategies, and Tasks

DOMINIK RIEDELBAUCH*, NICO HÖLLERICH*, AND DOMINIK HENRICH
Chair for Applied Computer Science III (Robotics and Embedded Systems), University of Bayreuth, 95440 Bayreuth, Germany

Corresponding author: Nico Höllerich (nico.hoellerich@uni-bayreuth.de)

This work was supported in part by the German Research Foundation (DFG) under Grant He2696/20 (FlexCobot) and Grant 491183248, and in part by the Open Access Publishing Fund of the University of Bayreuth.

*Dominik Riedelbauch and Nico Höllerich contributed equally to this work.

ABSTRACT Human-robot teaming receives an ever-increasing level of attention in research, development and industry. Novel approaches to task sharing in hybrid teams range from optimized schedules to intelligent cobot assistants with a high degree of autonomy. These approaches must prove their usefulness and benefits compared to manual work or full automation – particularly when it comes to assessing their potential for productive industrial use. This leverages demand for standardized, repeatable benchmarks to compare approaches and measure improvements in a reproducible way. Designing such benchmarks is challenging as numerous aspects, from safety considerations to human factors and team performance, must be considered. This survey seeks to contribute to the future development of benchmarks for the field of collaborative assembly, handling, and industrial cobot applications by giving a comprehensive overview of relevant metrics, evaluation strategies, and tasks for human-robot teams.

INDEX TERMS Human–robot interaction, intelligent robots, benchmarks, human factors, survey.

I. INTRODUCTION

Human-robot teaming has lately gained increased attention also in the context of industrial applications. Hybrid teams of humans and lightweight robots promise the symbiotic use of individual human and robot strengths in small and medium-sized enterprises (SMEs) [1]. Various planning methods for organizing joint task execution have been proposed: *Static teaming* approaches target the calculation of fixed schedules for mixed teams (e.g. [2], [3], [4], [5], [6]). By contrast, *dynamic teaming* methods emphasize situation-dependent co-working similar to human team coordination (e.g. [7], [8], [9], [10], [11], [12], [13]). These methods leverage robot reasoning and decision-making competencies to achieve dynamic online adaptation, e.g. regarding varying assembly sequences, product variety, worker availability etc. They integrate several complex techniques from different fields of cognitive robotics and artificial intelligence to make a robot communicate, sense, plan, and act together with humans in possibly unknown environments.

Independently of the teaming mode (static or dynamic), the introduction of robot co-workers in SMEs comes along with investment costs for suitable manipulators and sensory equipment – particularly cognitive robot systems as proposed in recent years must therefore be thoroughly tested regarding their usefulness to justify the investment. This leverages demand for structured benchmarks enabling the comparative assessment of different approaches, which is increased by the currently prevalent replication crisis in human-robot interaction (HRI) research [14]. Compared to prior benchmarks for intelligent robots (cf. e.g. [15], [16], [17]), establishing such benchmarks for *collaborative* robots is even more challenging. Metrics, strategies, and tasks must take human and robot into account: Firstly, human-robot teaming requires an evaluation along multiple scales in addition to task efficiency and effectiveness [18], e.g. human factors and occupational safety. Secondly, these scales are partly based on subjective metrics which can not directly be measured. And thirdly,

The associate editor coordinating the review of this manuscript and approving it for publication was Shaohua Wan.

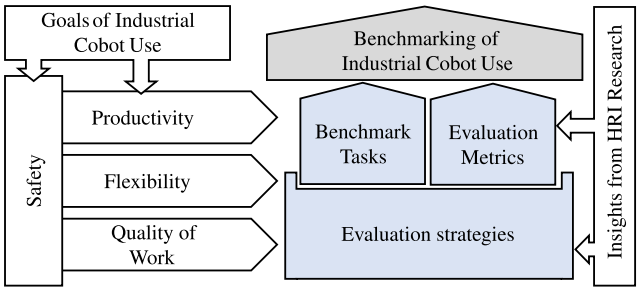

FIGURE 1. This paper addresses the key steps for benchmarking industrial cobot use: The process starts with the goals of cobot use under the constraint of ensuring safety. Evaluation strategies are then derived. Based on these strategies, concrete tasks and metrics are picked. Insights from HRI research influence metrics and strategies.

an additional challenge is raised by dynamic methods: Static approaches produce fixed human-robot workflows that are optimal according to some optimization criterion (e.g. physical ergonomics [5]). This yields an objective standard of comparison. By contrast, dynamic approaches are designed to adapt robot behaviour to different situations online. These situations emerge e.g. from human decisions or errors. Hence, teaming performance depends on the interplay of robot and human decisions, actions, and events that are not necessarily deterministic and fully foreseeable for robots. Benchmarking strategies for dynamic teams must therefore cover a range of situations which might occur during joint task execution to explore the average performance to expect from adaptive robot assistants.

Prior publications have already gathered sets of metrics for objective as well as subjective and cognitive aspects in the general field of HRI [19], [20], [21], [22], [23], [24]. In contrast to these surveys, this paper seeks to contribute more targeted insights for collaborative robotics and puts a stronger focus on the overall benchmarking process (Figure 1): After a short description of the literature review process (Section II), we will first reframe relevant metrics from HRI research with regard to typical goals of human-robot teaming in industrial settings (Section III). Strategies for collecting these metrics are then reviewed and discussed regarding their ease of use, reproducibility, and versatility (Section IV). Finally, we will give an overview of recently used tasks and model sets which might inspire future unified benchmark problems (Section V). To the best of our knowledge, this holistic view on hybrid teaming benchmarks has not yet been taken in literature.

II. METHODOLOGY

This survey is based on literature we gathered during our work in the field of human-robot teaming [25], [26], [27], [28], [29], [30], starting from 2017. We have expanded and completed this literature collection through an exploratory forward and backward search. For each item in our initial collection, we recursively screened the works referenced by the item and more recent works referencing the item. For the latter, we used the Google Scholar 'Cited by' functionality. The key criterion to include a publication was whether individual studies involved *task sharing* between human and robot agents according to the following nomenclature: A *task* or *process* $(T, \prec_T)$ is a pair of a set $T = \{\tau_1, \ldots, \tau_{|T|}\}$ of $|T|$ *subtasks* or *process steps* denoted τ_i $(i \in \{1, \ldots, |T|\})$ and a partial order $\prec_T$ that encodes "earlier-later"-relations (also called *precedence constraints*) between subtasks. Each subtask τ_i is assumed to be feasible for only one of the agents (human or robot) exclusively, for both agents alike, or for neither of them on his/her own in the case of timely synchronized collaboration. The problem we seek to benchmark is, hence, the coordinated division of subtasks among the members of a hybrid team. Accordingly, we will not cover benchmarking the performance of individual robot system components (perception, navigation, manipulation, etc.; see [23]) but rather consider measuring the effects of multi-agent teams as a whole. We will refer to the task definition when formalizing evaluation metrics in the next section.

III. EVALUATION METRICS

Metrics for measuring aspects of HRI have comprehensively been discussed [19], [22], [23], [24], [31], [32], [33], [34], [35]. In this section, we summarize this extensive body of work. We extracted metrics that can be numerically quantified and compared and put them into the context of industrial human-robot task sharing. These metrics are hereinafter clustered by the major goals that companies pursue when considering human-robot co-working for production. These are [1]

- *increasing productivity* by combining the strengths of humans and robots and sharing work in situations in which full automation of a process would be inefficient,
- *increasing flexibility* to reflect the trend towards small-batch production for mass customization without the long changeover times of classical automation systems,
- and *increasing job attractiveness* to improve the reputation of manufacturing work by reducing physical and mental stress and fostering innovative technologies.

In line with these goals, we have clustered metrics related to productivity as a profitability-oriented view on mixed teams (Section III-A), those measuring flexibility (Section III-B), and those suitable to quantify job quality (Section III-C). Seeking to achieve these goals can not be thought of without ensuring occupational safety in line with applicable laws and standards, e.g. defined by ISO/TS 15066 [36]. Therefore, we have also included metrics related to worker safety (Section III-D). Figure 2 summarizes the metrics considered hereinafter.

A. PRODUCTIVITY

Gaining productivity by leveraging individual human and robot strengths and division of work is a major motivation

Productivity	Efficiency	Cooperative Speed-Up [10]
		Relative Helpfulness [37]
		Capability Indicators, e.g. [38]
		Robot Error Rate
	Teaming Fluency	Human/Robot Idle Time [34]
		Concurrent Activity [34]
		Robot Participation Rate [10]
Flexibility	Task Flexibility	Normalized Teaching Time [39]
		Teaching-to-Use Time Rate [39]
	Teaming Flexibility	Levels of Robot Autonomy [40]
		Intervention Rate [41]
		Neglect Tolerance [35]
Job Quality	Phys. Ergonomics, e.g.	SI [42]
		REBA [43]
		RULA [44]
	Cog. Ergonomics, e.g.	NASA-TLX [45]
		SUS [46]
		STAI [47]
		Team Fluency [34]
Safety	Physical Safety	Separation Distance [33]
		Contact Force/Pressure
	Psychological Safety	Custom Questionnaires, e.g. [48]–[51]
		Anxiety by Physiological Signals [52]
		Posture [53]
Combined	Measure	Throughput Rate per Unit of Work Effort Time [54]
	Framework	UX [55]

FIGURE 2. Overview of relevant metrics for assessing human-robot teaming methods with regard to common goals in industrial scenarios.

for introducing human-robot teams. We assume that workers are generally skilled and that robots are also equipped with appropriate manipulation and perception skills to work effectively towards the correct completion of the shared task. Based on these assumptions, the first cluster of metrics targets the effects of hybrid teams on goal achievement with regard to **efficiency**. Definitions of "efficient goal achievement" are, of course, strongly application-dependent. For the manufacturing domain, we propose to consider the following aspects: The *overall time to completion D* is the time needed to finish a task [23], e.g. assembling one product instance. This time span is also referred to as *makespan* [5] or *cycle time* [56] in the terminology of production key performance indicators. Let D_{H} denote the duration of the purely manual process as carried out by a human worker H. Concrete values can then be estimated with standard motion time systems (e.g.Methods-Time Measurement (MTM) [57]). Further, let $D_{\mathrm{H/R}}$ denote the duration of the same task when partly automated by a human-robot team. We then propose defining the *cooperative speed-up* $\mathcal{S}_{\mathrm{H/R}}$ [10] as

$$\mathcal{S}_{\mathrm{H/R}} = \frac{D_{\mathrm{H}}}{D_{\mathrm{H/R}}}. \tag{1}$$

A desirable cobot system must achieve high speed-up values across different tasks – reducing the time to completion carries over to production costs induced by human labour as well as robot energy consumption.

Reducing task durations by work sharing is not only a matter of economic considerations but also from a perspective of technology acceptance: Expectations of being relieved from parts of a task have been found to influence workers' attitudes towards robots positively [58] – this brings us to the notion of helpful robots "trying to play a positive role [for humans] with the task at hand" [37]. Therefore, a robot co-worker should generally be perceived as a helpful partner. Freedman et al. have lately proposed *relative helpfulness* $\mathcal{H}_{\mathrm{R}}$ as a quantitative metric that relates the decrease in generic costs to achieve some goal by human-robot teaming to the cost of human-only task execution [37]. With this definition, $\mathcal{H}_{\mathrm{R}}$ is directly related to cooperative speed-up (Equation 1) if we take time to completion as a measure for task costs:

$$\mathcal{H}_{\mathrm{R}} = \frac{D_{\mathrm{H}} - D_{\mathrm{H/R}}}{D_{\mathrm{H}}} = 1 - \frac{1}{\mathcal{S}_{\mathrm{H/R}}} \tag{2}$$

Although dual to cooperative speed-up, relative helpfulness provides a more worker-centred quantitative view on the overall relief induced by robot teammates.

The alignment of subtask allocation with individual agent capabilities is another central aspect of efficiency in human-robot teaming. E.g., humans are still ascribed superior sensorimotor abilities, whereas robots exceed in precision [1]. It would therefore be inefficient to allocate a strongly dexterous subtask to a robot while the worker is assigned process steps to place small workpieces with high precision in the meantime – in this example, each subtask would likely be performed far more quickly by the other agent in line with respective capabilities. Structured methods to determine so-called *capability indicators* have been proposed early [38]. These indicators rate to what extent subtasks τ are suitable for execution by either human or robot. To this end, they typically condense several criteria into real-valued scores $c_{\mathrm{H}}(\tau)$ for humans and $c_{\mathrm{R}}(\tau)$ for robots, with higher indicator values meaning better suitability of the corresponding agent. Static planning approaches use capability indicators as optimization criteria before task execution (e.g. [6], [38]). Yet, accumulated capability indicators of subtasks assigned to each agent can analogously be used to evaluate the quality of decisions for dynamic teaming approaches after task execution.

Agile robots can make mistakes [59], especially when humans are around and unpredictably modify parts in the workspace. Consider e.g. manipulation failure [23], [59] or erroneous robot task allocation decisions colliding with human actions due to incomplete knowledge of the task progress [10]: any human or robot error requires time to resolve. This timespan is lost from the productivity point of view. The number of errors should therefore be considered an important metric for mixed teams [19]. As we are focussing on the impact on productivity here, the concrete error source is of no importance. Consequently, we can say that trying

to execute a subtask can either succeed or fail. Overall, the robot will successfully complete $N_{\mathrm{R}}^{\mathrm{success}} \leq |T|$ of all $|T|$ subtasks. To this end, it will issue a cumulated number of $N_{\mathrm{R}}^{\mathrm{attempt}} \geq N_{\mathrm{R}}^{\mathrm{success}}$ attempts to handle subtasks. The number of attempts may exceed the number of completed subtasks, as subtasks that failed may be retried several times. We can then define the *robot error rate* $\mathcal{E}_{\mathrm{R}}$ as a relative metric:

$$\mathcal{E}_{\mathrm{R}} = 1 - \frac{N_{\mathrm{R}}^{\mathrm{success}}}{N_{\mathrm{R}}^{\mathrm{attempt}}} \tag{3}$$

This metric can analogously be used to measure human error.

Teaming fluency, as discussed by Hoffman [34], helps to analyse the timing aspect during shared task execution beyond makespans. The corresponding metrics are connected with efficiency and productivity as follows: *Human Idle Time* $D_{\mathrm{H}}^{\mathrm{idle}}$ accumulates timespans during which the human was willing to contribute to task progress but was delayed, e.g. while waiting for the robot to fulfil an "earlier-later"-constraint. Similarly, phases of *Robot Idle Time* $D_{\mathrm{R}}^{\mathrm{idle}}$ can emerge while waiting for human input or action. Both sorts of idle time indicate pure utilization of production resources and should be minimized from the perspective of productivity. By contrast, measuring long timespans D^{coop} of *Concurrent Activity*, during which human and robot are equally active, and a small deviation $D_{\mathrm{H/R}} - D^{\mathrm{coop}}$ between concurrent work and overall task duration indicates a successful division of work and efficient use of available agent capacities. Normalizing the aforementioned measures by the task duration $D_{\mathrm{H/R}}$ renders them comparable across different tasks and robot systems. Hoffman [34] furthermore proposed to track the *functional delay* as the time between an agent's action and the beginning of her/his partner's action. This metric is particularly relevant for approaches where the agents take turns one after another or in the context of collaborative interaction (e.g. when waiting for a part to be handed over); however, this metric is redundant from a productivity-centred point of view, as idle time includes such timespans.

Concurrent activity D^{coop} provides a basic understanding of resource utilization. Still, this metric does not necessarily indicate a proper division of labour amongst agents. Compared to humans, robots may work slowly, e.g. due to safety considerations or limited dexterity. Teamwork will then not speed up tasks significantly. We have previously used the *robot participation rate* $\mathcal{P}_{\mathrm{R}}$ [10] to capture this aspect. This metric relates the number of subtasks N_{R} handled by the robot to the overall number of subtasks $|T|$:

$$\mathcal{P}_{\mathrm{R}} = \frac{N_{\mathrm{R}}}{|T|}. \tag{4}$$

Assuming subtasks of similar durations (as is the case in e.g. pick-and-place tasks within the bounded workspace of typical cobots), a competitive robot co-worker should achieve $\mathcal{P}_{\mathrm{R}}$ values close to the fraction of robot to human working speed. Otherwise, the system is not performing as a partner up to its potential. This may be caused by the performance of perception and planning components or by coordination schemes with an overhead of interaction effort – such issues would also be indicated by negative helpfulness scores ($\mathcal{H}_{\mathrm{R}} < 0$) and speed-up values $\mathcal{S}_{\mathrm{H/R}} < 1$. When interpreting $\mathcal{P}_{\mathrm{R}}$, one must consider that its value is capped to the percentage of subtasks the robot can contribute to – if subtasks exist that only humans are capable of (Section I).

B. FLEXIBILITY

Low changeover times in small-batch production are a major goal of industrial human-robot teaming [1] that we refer to by **task flexibility**. Programming has been identified as the most time-consuming activity in this context [60], and we can hence assume this step to also constitute the major part of overall changeover times in robot co-working systems. To quantify this effort, Marvel et al. proposed using the *programming time* as a metric for multi-robot teams. As there are various programming approaches used for instructing collaborative robots (e.g. visual programming with skills [10] or learning from demonstration [61]), the more general term *teaching time* D_{teach} will be used hereinafter. It is important to notice that this term does not necessarily have to refer to the teaching of robots only: safe and efficient co-working is supported by well-trained employees [62], and qualification times can thus additionally be taken into account when measuring D_{teach}. This absolute metric is hard to compare across tasks. We, therefore, propose the *normalized teaching time* $\widetilde{\mathcal{D}}_{\mathrm{teach}}$: Assume $\mathrm{TA}(\tau)$ to break down a subtask $\tau \in T$ of some task $(T, \prec_{\mathrm{T}})$ into a set of work items with equally short durations using standard task analysis (TA) techniques (e.g. MTM [57]). The normalized teaching time is then given by

$$\widetilde{\mathcal{D}}_{\mathrm{teach}} = \frac{D_{\mathrm{teach}}}{\Sigma_{\tau \in T} |\mathrm{TA}(\tau)|}. \tag{5}$$

According to this definition, low teaching times per work item ($\widetilde{\mathcal{D}}_{\mathrm{teach}} \to 0$) indicate a robot teammate that can quickly be adapted when partial automation of a new task is desired.

Ongoing operational costs are a key concern when introducing hybrid teams [1]. Even if low $\widetilde{\mathcal{D}}_{\mathrm{teach}}$ scores are achieved, the absolute effort for commissioning may still prevent profitable system operation. Therefore, the duration of teaching a new task must also be put in relation to the subsequent use times [39]. This is achieved by the teaching-to-use time rate $\mathcal{T}$ with

$$\mathcal{T} = \frac{D_{\mathrm{teach}}}{N_{\mathrm{lot}} \cdot D_{\mathrm{H/R}}}, \tag{6}$$

where $N_{\mathrm{lot}} \in \mathbb{N}$ denotes the lot size to produce. In line with the effects of automated mass production, $\mathcal{T}$ tends towards zero with increasing lot sizes. When comparing two human-robot teaming approaches for small-scale partial automation, lower $\mathcal{T}$ scores indicate better task flexibility. Teaching times should, in any case, not exceed the gain in productivity as expressed by a decrease $D_{\mathrm{H}} - D_{\mathrm{H/R}}$ in production times per task execution (Section III-A). This interplay between productivity and task flexibility metrics can be expressed with an alternative formulation $\mathcal{H}'$ of relative helpfulness

(Equation 2): As this metric is defined for generic costs [37], the lot size and teaching time can be integrated by adding the contribution $\frac{D_{\text{teach}}}{N_{\text{lot}}}$ of teaching per production cycle to the expected human-robot time to completion $D_{\text{H/R}}$:

$$\mathcal{H}' = 1 - \frac{\frac{D_{\text{teach}}}{N_{\text{lot}}} + D_{\text{H/R}}}{D_{\text{H}}} \tag{7}$$

In analogy to Equation 2, high scores for $\mathcal{H}'$ indicate a helpful robot with regard to a given task. When we assume that teamwork speeds up tasks ($D_{\text{H/R}} < D_{\text{H}}$, or respectively $\mathcal{H} > 0$), negative values of $\mathcal{H}'$ carry the additional information that teaching efforts exceed the achieved gain in productivity.

In addition to task flexibility, particularly dynamic teaming methods are also directed towards **teaming flexibility**, i.e. towards adapting to variance during the joint execution process. A need for such adaptations can e.g.arise in situations when a worker leaves temporarily, e.g. at shift changes, or when handling a more important intermediate task is necessary [10], [13]. Further aspects can be derived from considerations on robot agility in general [59]: To what extent can a robot handle (self- or human-induced) failure in the process? Can it adapt to environmental changes, e.g. a tool moved to a different position by workers? All these aspects of teaming flexibility can be covered by classifying the autonomy level of a robot co-worker. To this end, early definitions of *Levels of Automation* (LOA) for the division of work among humans and any sort of machines in general (e.g. [63], [64]) have inspired more specialized taxonomic frameworks for HRI. In particular, the *Levels Of Robot Autonomy* (LORA) proposed by Beer et al. [40] are suited for evaluating human-robot teaming [18]. Respective frameworks enable a categorization of the overall system, e.g. in the case of LORA by evaluating the function allocation to either human or robot for the basic actions of sensing, planning, and acting [40].

For the task-sharing scenarios under consideration in this survey, answers to the question of whether certain functions are allocated to the robot are not necessarily binary for a given system – they may depend on the concrete task that a mixed team works on: Consider e.g. a stationary robot manipulator with a camera attached to the robot hand as the only sensor. According to the LORA taxonomy, a task-sharing system using this sensory setup will certainly implement sensing capabilities for object detection in general. Still, the degree of autonomy will depend on the concrete task design. If parts are out of perception range, support by the partner is required to gather information on these objects. Similarly, the capability to act on parts may vary within a continuum depending on part locations and robot reach. A quantitative placement along this continuum of shared control in which humans, as well as robots, are accountable for a (possibly sliding) amount of allocated functions [41] can be achieved by measuring the amount of intervention or using the related neglect tolerance metric [40]:

The *amount of intervention* is defined as the fraction of time during which a human controls the robot [41]. Yanco and Drury have proposed it in the context of mobile robots able to navigate with varying amounts of control by a human supervisor – this is opposed to our task-sharing setting in which the human task is not supervision but equally contributing to a shared goal. We can still reframe this metric accordingly: Human intervention can be assumed to occur as punctual events of short duration in the context of productive robot co-workers, e.g. providing pieces of information, supporting collaborative subtasks, or especially helping out in case of robot failure. The amount of intervention can then be measured implicitly via other metrics, e.g.the interaction effort (Section III-C) or the robot error rate (Equation 3).

Yanco and Drury have further defined the autonomy level as ''the percentage of time that the robot is carrying out its task on its own'', complementary to the amount of intervention [41]. Similarly, *neglect tolerance* has been introduced by Olsen and Goodrich [35]. Neglect tolerance quantifies the timespan that a system can work without human attention or intervention on a level of effectiveness that is considered sufficient (so-called *neglect time*). Although originally motivated by man-machine interface design for remote robot operation, neglect tolerance has semantics well-suited for industrial co-working: High neglect tolerance expresses a productive, failure-proof robot even in situations when humans are not available or not willing to interact. Consider e.g.a cobot which needs humans to explicitly confirm each subtask they completed (e.g. [65], [66]). Systems relying on this coordination scheme will stop working if the input on task progress is no longer provided. They have low neglect tolerance. By contrast, approaches with implicit coordination by action or world state observation (e.g. [10], [11]) are more neglect tolerant. Human attention is here only needed when a subtask requires the collaboration of both agents simultaneously. For precise measurements of neglect tolerance, Olsen et al.have proposed to observe the mean neglect time $\bar{D}_{\text{neg}}$ between two subsequent user interactions by observing the points in time when users actually intervene, or effectiveness drops below the acceptable threshold [35]. For our nomenclature regarding task sharing (Section I), we can e.g.say that a robot is sufficiently effective as long as it productively works on one of the subtasks and that it is no longer effective when entering an idle time phase (Section III-A). As with participation rates (Equation 4), neglect tolerance must be carefully interpreted in the context of the task at hand. For the same robot system, a task which is dominated by subtasks that only humans are capable of might lead to low neglect tolerance. In contrast, tasks with a high potential for parallel work will yield high neglect times.

C. JOB QUALITY

In addition to productivity and flexibility, another goal of cobot use is to enhance human job quality. There are two major research fields which offer metrics to quantify working conditions: *physical ergonomics* assesses loads impacting the

human body to prevent disorders of muscles, nerves, and joints (Section III-C1). By contrast, *cognitive ergonomics* aims at mental health and perceived comfort (Section III-C2). We will hereinafter review relevant metrics from both areas.

1) PHYSICAL ERGONOMICS

There are several established tools for assessing the physical ergonomics of a workplace: frameworks such as the *Strain Index* (SI) [42], *Revised Strain Index* (RSI) [67], *Rapid Entire Body Assessment* (REBA) [43], *Rapid Upper Limb Assessment* (RULA) [44], *European Assembly Worksheet* (EAWS) [68], or the *Washington Industrial Safety and Health Act* (WISHA) [69] have been used as indicators to evaluate human-robot workflows with regard to human physical strain (e.g. [2], [3], [5], [18], [70], [71], [72], [73]). These frameworks condense the information on the human posture (encoded by limb angles, the weight of handled parts etc.) into a single numerical score. Typically, lower values indicate ergonomically favourable situations – accordingly, unfavourable subtasks can be shifted to robot co-workers to reduce the physical load by capability-based task allocation. On top of the indicators for single process steps or motions, attempts have been made to cumulatively rate the effects of sequences of actions, ultimately taking muscle fatigue into consideration [71]. Aside from the ergonomics scores based on observing limb angles with the aforementioned frameworks, *electromyography* (EMG) records of the electrical signals transmitted to muscles can be used to estimate muscle activity and fatigue online [74], [75], [76], [77]. We refer the reader to further literature that discusses measurement methods for physical risk assessment in depth [31], [78].

2) COGNITIVE ERGONOMICS

Improving cognitive ergonomics can influence productivity and product quality positively [79], [80]. Hence, the benefits of task sharing and enhanced physical ergonomics should not be compensated by the negative impacts of cobots on cognitive ergonomics, e.g. increased stress or fatigue as a consequence of high mental workload [81], [82] or of resistance to cooperating with the robot [83]. Contrasting to the above productivity, flexibility, and physical ergonomics metrics, assessing concepts related to human factors is more challenging as they concern subjective human impressions during the teaming processes. These are investigated with two predominant strategies, which the remainder of this section puts emphasis on:

- **Questionnaires** are the most common tool for participant self-reporting in human factors analysis. They are frequently composed of questions to rate one's impression on a 5- or 7-point Likert scale and have been applied to measuring a broad range of specific aspects.
- **Physiological Measurements** can be used to monitor respiratory rates, *heart activity* (electrocardiography, ECG), *brain activity* (electroencephalography, EEG), *eye activity* (electrooculography, EOG), or the *electrodermal activity* (EDA; also known as *galvanic skin response*).

Aside from these methods, there are further, less frequent strategies which we name for the sake of completeness: (i) *Direct Input Devices*, such as joysticks or sliders, allow subjects to input values continuously. The input data can then e.g.be used to quantify emotions in terms of valence and arousal [84], [85]. (ii) *Behavioural Assessment* relies on video recordings of subjects to categorize their behaviour after the actual experiment [27], [86], [87], [88]. (iii) *Computational Models* relate quantitative values to human factor concepts, e.g.trust to objective metrics [89], [90], [91], [92], cognitive workload to physiological signals [93], [94], stress to skin temperature [95], or anxiety to facial expressions [96].

These methods are used to collect data on a wide variety of human factors. To identify concepts previously raised in the context of human-robot task sharing, we reviewed prior surveys of Baltrusch et al. [97], Hopko et al. [98], Lorenzini et al. [31], Nelles et al. [21], Rubagotti et al. [52] and Wurhofer et al. [99]. We unified their terminology and extended them towards recent publications. This gave us the below list of concepts with prominent examples of applied measurement instruments (see Tables 1 and 2 for a comprehensive list of questionnaires and physiological measures used with these concepts):

Cognitive Workload is a ''multidimensional concept that consists of four components: 1) task complexity; 2) mental workload; 3) performance; and 4) depletion factors (e.g.stress, fatigue, motivation) '' [143]. Reaching the limits of one's cognitive capacities can lead to stress and anxiety [82]. In the long run, high mental workload causes fatigue, increases error rates, and hence decreases performance. The well-known *NASA Task Load Index* (NASA-TLX) [45] puts emphasis on the perceived physical and cognitive workload when using a system. Helton et al. have used an extension of the NASA-TLX towards perceived teaming workload (e.g. in terms of perceived effort for coordination and communication, team support etc.) [100]. If only cognitive workload is relevant, the *Subjective Workload Assessment Technique* (SWAT) [102] is a validated alternative. When only mental effort is relevant, the *Rating Scale Mental Effort* (RSME) [103] is another established measurement instrument. It gives users more guidance by providing descriptions at certain scale levels [150]. These questionnaires can be complemented with physiological signals related to cognitive load (Table 2).

Affect refers to the experience of feelings, emotions, or mood [151]. Negative emotions towards the robot and the interaction with it may degrade trust and acceptance. Frequently investigated sub-aspects of affect are anxiety, frustration, emotional stress and (dis)comfort. Prominent scales with a focus on measuring these sub-aspects are the *State-Trait Anxiety Inventory* (STAI) [47], the *Positive and Negative Affect Schedule* (PANAS) [105], and the *Negative Attitudes Towards Robots* scale (NARS) [106], [107].

TABLE 1. Aspects of job quality and psychological safety covered by questionnaires. Half circles indicate that only one question targets the concept or an aspect of it. Full circles denote multiple questions.

	Human Factors						Teamwork		
	Cognitive Workload	Affect	Psychological Safety	Satisfaction	Subjective Performance	Acceptance	Personality	Interaction Quality	Trust
Established and validated scales from Human Factors Research									
NASA-TLX [45]	◑	○	○	◑	◑	○	○	○	○
NASA-TLX + Team Workload [100]	◑	○	○	◑	◑	○	○	●	○
SUS [46]	○	○	○	●	○	●	○	○	○
UMUX [101]	○	○	○	●	◑	◑	○	○	○
SWAT [102]	●	○	○	○	○	○	○	○	○
RSME [103]	◑	○	○	○	○	○	○	○	○
Godspeed [104]	○	○	◑	○	○	○	●	○	○
STAI [47]	○	●	○	○	○	○	○	○	○
PANAS [105]	○	●	○	○	○	○	○	○	○
NARS [106], [107]	○	●	○	○	○	○	○	●	○
UTAUT [108]	●	●	○	●	●	●	○	○	○
Variations of the human-robot teaming fluency questionnaire									
Hoffman [34]	○	●	○	◑	●	○	●	●	●
Gombolay et al. [109], [110]	○	●	○	○	●	◑	●	◑	●
Dragan et al. [48]	○	◑	●	◑	●	○	●	●	●
Nikolaidis et al. [61]	○	○	○	○	◑	○	●	○	●
Unhelkar et al. [51]	◑	●	●	○	●	○	○	●	●
Paliga and Pollak [111]	○	○	○	○	●	○	○	◑	◑
Validated scales for human-robot collaboration									
Charalambous et al. [50]	○	●	●	●	○	○	○	○	●
Inoue et al. [49]	○	●	●	○	○	○	○	○	○
Bröhl et al. [112], [113]	○	●	◑	◑	●	●	○	○	○
Schaefer [114]	○	○	○	●	●	○	●	●	●
Further questionnaires gathered from individual studies									
Baraglia et al. [87]	○	○	○	◑	●	○	●	●	○
Etzi et al. [115]	◑	●	◑	◑	○	○	○	◑	○
Fratczak et al. [53]	◑	○	○	◑	●	○	●	○	○
Chao and Thomaz [116]	◑	◑	○	●	●	○	○	●	◑
Palmarini et al. [117]	○	○	●	○	○	○	○	○	○
Lasota and Shah [118]	○	○	●	○	○	○	○	●	●

TABLE 2. Overview of studies which target human-robot cooperation and use physiological measures for human factors analysis. References in grey indicate that the study did not find statistically significant differences between conditions with respect to the concept.

	Trust	Cognitive Workload	Anxiety	Situational Awareness
ECG		[119]–[124] [115], [125]	[53], [84], [122], [123], [126]–[129] [115], [125], [130]	
EEG		[93], [131]		
EOG and gaze		[120], [132]	[133] [134]	[135], [136]
EDA		[137] [115], [120]	[84], [126]–[130], [134], [137]–[140] [133]	
Respiratory Rate			[53], [123], [141]	
Posture	[86]	[120]	[53] [134]	
Surveys	[142]	[143]–[146]	[147], [148]	[149]

Satisfaction: Satisfaction is the "extent to which the user's physical, cognitive and emotional responses that result from the use of a system (...) meet the user's needs and expectations" [152]. Important factors influencing satisfaction are feature consistency, robot support [100], contentment with the interaction [99], self-efficacy [108], trust, and pleasure/frustration [152]. Satisfaction is (besides efficiency and effectiveness) a core dimension of usability [152]. The most prominent usability evaluation method is the *System Usability Scale* (SUS) [46]. It accurately captures the usability and learnability of a system [153] while only moderately correlating with task performance [154]. A compact variant is the *Usability Metric for User Experience* (UMUX) [101].

Subjective Performance complements the objective performance metrics (Section III-A). Respective questionnaires capture how subjects perceive overall efficiency, effectiveness, and output quality, as well as the individual contribution of human and robot to the task. E.g., objective productivity metrics as the concurrent activity have counterparts capturing human subjective ratings of teaming performance in variations of the *fluency questionnaire* [34].

Acceptance encompasses a person's attitude and behaviour towards a robot [58], and behavioural acceptance may range from commitment to refusal. Early models for measuring the acceptability of automation are the *Technology Acceptance Model* (TAM) [155] and the *Unified Theory of Acceptance and Use of Technology* (UTAUT) [108]. When using more recent extensions of these models, as proposed by Venkatesh et al. [156], [157], it must be taken into account that interaction with technology can be mandatory rather than voluntary in industrial settings [158], [159]. A better-suited variant, particularly for human-robot cooperation, has been constructed and validated by Bröhl et al. with *ease of use*, *usefulness*, and *intention to use* as core factors [113].

Personality refers to the personality traits that humans attribute to the robot based on its behaviour. Examples are likeability, intelligence, appreciation, respect, cooperativeness or legibility of behaviour. The *Godspeed* questionnaire is a well-known example of capturing these aspects [104].

Interaction Quality refers to the perceived fluency and naturalness during joint task execution. Coordination, communication, and time-sharing demands [100], as well as the experienced teaming and waiting times, are important factors influencing this concept. Similar to subjective performance, some aspects of interaction quality can objectively be measured, e.g.with productivity-related fluency metrics such as human idle times, functional delays [34], or the interaction effort [35]. A questionnaire for the *subjective fluency* of HRI has been introduced by Hoffman et al. [160] and modified for several studies [48], [61], [109], [161]. Paliga et al.have refined the concept into human-oriented, robot-oriented, and team-oriented components [111]. Importantly, subjective fluency scores do not necessarily correlate with corresponding objective measures: a cumbersome interaction which objectively increases efficiency can feel less fluent than a natural but less performant interaction [34].

Trust is a multidimensional concept which emerges from the interaction of two partners. Wurhofer et al. define it as "the extent to which the user feels confident that the system will behave as intended" [99]. An appropriate level of trust forms a keystone towards efficient teamwork [89]: Both overtrust (i.e.users *over*estimating robot capabilities) and distrust (i.e.users *under*estimating a robot and, hence, intervening too frequently) can degrade the overall team performance [162]. Established measurement instruments for trust in automation exist [163], [164], but these are not directly usable for HRI [165], where robots can act autonomously and humans may play the role of a teammate [166]. In consequence, modelling techniques (see e.g.the surveys of Khavas et al. [167], Hancock et al. [168]) and items in several questionnaires (Table 1) have been proposed to specifically target trust in HRI. Yet results must be cautiously interpreted since subjects tend to put more trust in robots when experiments are conducted in controlled lab settings [48].

Situational Awareness (SA) is "the perception of the elements in the environment within a volume of time and space, comprehension of their meaning and the projection of their status in the near future" [169]. High SA can thus help to predict and prevent mistakes. We found three major strategies to evaluate SA: (i) *Freeze-probes* freeze the task at some point in time and ask the user questions about his/her understanding of the situation [51], [110]. (ii) *Attention* can be measured by observing for how long subjects' gaze is directed towards certain regions, either with an eye-tracking system [96], [136] or with questionnaires [53]. (iii) *Intervention* can be observed by intentionally creating exceptional situations and giving users a small time window to understand and react [136].

D. SAFETY

Keeping humans safe at any time during HRI is a primary obligation for ethical reasons in general and due to laws and regulations in manufacturing environments in particular. Two sorts of safety are commonly distinguished [20]: *Physical safety* is concerned with preventing unintended, forceful human-robot contact, which might injure the human body. Even if a robot is technically capable of stopping in time to avoid such harm, too fast motions close to a human may still cause discomfort [85] – this sort of psychological harm as also induced by stress, anxiety, or the violation of social norms renders an additional consideration of *psychological safety* necessary. The achievement of safety can be measured with the following metrics:

Regarding **physical safety**, the International Organization for Standardization defines the *protective separation distance* and *force limits*, which may not be exceeded in case of direct contact [36]. The protective separation distance is defined as the minimum distance that lets a robot stop before colliding with a human based on speed, reaction times and positioning

uncertainties. Accordingly, comparing the current distance (and the currently exerted force in case of wanted human-robot contacts) with the protective separation distance can be used as a metric to judge the obtained level of physical safety. This comparison requires estimating the human body pose, e.g.by inertial measurement units (IMU) [170], [171], [172], [173], by colour segmentation [174], by skeleton tracking in camera images [175], [176], with data gloves [177], [178], or with dedicated markers attached to the body and captured by a surrounding detection system [179], [180], [181], [182], [183]. An alternative perspective on separation distance is the *time to collision* [184]. We direct the reader to Kumar et al. [33] for an in-depth discussion of speed and separation monitoring and a list of measures.

According to Rubagotti et al., the core concepts related to **psychological safety** are trust, comfort, stress, fear, anxiety, and surprise. We have already encountered these aspects in the context of cognitive ergonomics (Section III-C), where we subsumed a part of them (comfort, stress, and emotions like fear and surprise) under the notion of affect. As a consequence of this similarity, perceived psychological safety is often evaluated with questionnaire items in conjunction with these aspects (e.g. [48], [51], [104], see Table 1), items explicitly targeting safety (e.g. [49], [50], [112]), or, via physiological measurements in the case of anxiety (Table 2).

E. COMBINED METRICS

A benchmark to comprehensively evaluate human-robot teaming must ideally cover all relevant goals. By combining complementary metrics, one better understands a system's overall impact on workers. We found two ways in literature to achieve this: (i) Several goals can be combined into a single numerical indicator. Zhang et al. [54] have proposed the *Throughput Rate per Unit of Work Effort Time*. It combines productivity and physical ergonomics by putting throughput rates and the Strain Index into relation. Rephrased in the formalization nomenclature of this paper, this measure is defined by

$$\mathcal{C} = \frac{1/D_{\mathrm{H/R}}}{D_{\mathrm{H/R}} \cdot \mathrm{SI}} \tag{8}$$

with throughput rate $1/D_{H/R}$ [56] and the overall Strain Index [42] SI accumulated across all subtasks. (ii) Alternatively, several measurement instruments can be combined into more complex evaluation frameworks. Following this strategy, the framework of Gervasi et al. [18] embeds established metrics from Sections III-A to III-D (e.g.Levels of Robot Autonomy, NASA-TLX, EAWS, SUS) into a higher-level rating scheme. Similarly, Wallström and Lindblom [185] have proposed to combine measures for productivity (e.g.effectiveness and efficiency) and for job quality (e.g.trust and safety) into an HRI design process inspired by user experience (UX) design goals. An important feature of the latter method is that it does not require large sample sizes – a convenience sample is often sufficient, hence rendering the framework well-suited for early prototype design [185].

F. DISCUSSION

Productivity, flexibility, job quality and safety are the main goals for evaluating industrial human-robot applications. Metrics to measure productivity (Section III-A) and flexibility (Section III-B) are often objective and easily measurable by an observing experimenter (e.g.task completion time, idle time, teaching time). There are also objective metrics to quantify aspects of job quality (e.g.physical ergonomics frameworks, Section III-C) and safety (e.g.separation distances, Section III-D), some of them requiring more sophisticated measurements of physiological signals (e.g.anxiety). However, particularly job quality is strongly linked to human factors, which are predominantly evaluated with users' self-reports based on questionnaires. Physiological signals and questionnaire-based metrics come along with specific challenges, as discussed below – more general guidelines for the design and conduct of HRI studies to gather these metrics are outlined in Section IV-E.

Particularly questionnaires need to prove their reliability and validity, i.e.it must be shown that there is a "correlation between respondent's scores and the true level of the concept being measured" [186] and that there is a high "degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses" [187]. Long-term established and well-known standard questionnaires often have a broad range of literature showing under which circumstances reliability and validity can be assumed (e.g.SUS [46], NASA-TLX [45]). Whenever using modified or self-designed questionnaires, researchers should consider newly proving the reliability and validity as a part of their evaluation. This can e.g.be achieved by adhering to the structured design process for measuring new concepts as proposed by Rueben et al. [186] (Figure 3). A popular way to prove reliability is to investigate the internal consistency via the *tau-equivalent reliability* (formally known as *Cronbach's alpha*) – see Cho et al.for detailed guidelines [188]. In their work, Rueben et al.also discuss indicators related to validity in the process of creating and adapting questionnaires. Despite following these guidelines, the response process can still undermine the validity of the measurement instrument. Different sorts of biases need to be considered when interpreting results, e.g.a bias towards the extremes of a scale, the social desirability bias [189], or subjective biases resulting from one's personal (dis)likes [190]. Rosenthal et al. [191] give detailed insights into study design for behavioural assessment.

Compared to questionnaire-based self-reporting, physiological signals promise less subjective bias and higher sampling rates [121]. In turn, two other challenges arise: (i) According measurements of the underlying physical property must be obtained. This can partly be achieved with consumer devices, such as smartwatches or chest-strap sensors (e.g.for measuring the heart rate), but often requires

FIGURE 3. Process of measuring a new concept in an HRI experiment [186].

specialized medical devices (e.g.EEG). For the latter, high deployment costs, limitations of the human movement space, accuracy issues, and the problem of choosing an appropriate measurement horizon and baseline measurement must be considered [192], [193]. (ii) The measured values must relate to the concept under investigation, i.e.the question of reliability and validity arises for questionnaires and physiological measures alike. Finding a fitting measure and categorization is still challenging: The current literature is e.g.still inconclusive regarding the best-suited combination of physiological signals for measuring cognitive workload [31], [143], [144], [145], [146] when physical workload and anxiety towards the robot might change as well during the same experiment. Models developed so far might hence still be insufficient to match the results to be obtained with self-reports [115], [144], [145], [146], or insufficient to discriminate different conditions (grey references in Table 2).

A comprehensive evaluation needs more than one metric to cover all concepts relevant to human-robot teaming (Section III-E). However, using multiple measures and evaluation methods for a single concept within the same study should also be considered. This can help to reduce measurement errors and leads to more valid, reliable, and conflict-free results with regard to some concept [194]. The strategy of using three or more measures is generally called *triangulation* (see e.g. [195] for a general in-depth discussion). HRI studies have e.g.combined questionnaires with objective physiological measures [53], [196], [197] and/or task performance metrics [115], [198].

IV. EVALUATION STRATEGIES

Insights on human-robot teaming can be gathered by applying different evaluation strategies to a given system. We will discuss these hereinafter: Starting with considerations on research demonstrators in Section IV-A, Section IV-B gathers recent works on user study design. Section IV-C then elaborates on simulation-based evaluation strategies where HRI are observed (partly or fully) in virtual spaces. We conclude with analytical performance models in Section IV-D and discuss the properties and interrelations of individual methods in Section IV-E.

A. RESEARCH DEMONSTRATORS

Demonstrators play a vital role in research and development processes. Moultrie has formulated a comprehensive model that differentiates between several types of demonstrators according to their purpose [199]. For basic research on human-robot teams, *physical prototypes* are important artefacts for advancements under laboratory conditions [200]. Yet, partially or fully *virtual prototypes* of hybrid workstations may also be built with modern simulation and virtual reality techniques (e.g. [201], [202], [203]). Demonstrators can moreover serve as 'boundary objects' between individuals with different expertise (e.g. academic researchers and stakeholders from industry) [199] when approaching higher technology readiness levels. They are then used as common ground for structured interdisciplinary dialogue and participative design steps, e.g. when jointly investigating human-robot safety solutions [204]. Demonstrators are themselves an evaluation strategy as any proof-of-concept implementation directly measures 'feasibility' as a binary metric. This is usually achieved by showcasing concrete use-cases (e.g. [205]) or by reporting on the typical sequence of actions when interacting with a cobot (e.g. [206], [207], [208]).

B. HUMAN-PARTICIPANT STUDIES

Based on a prototype system, human-participant studies (often also called *user studies*) are frequently used experimental tools for evaluating HRI [209]. They enable users to work with a system so that researchers can observe and gather measures (cf. Section III) as required for judging and comparing the performance of different approaches. Human-participant studies will, hence, certainly be part of future benchmarking protocols. This section is intended to introduce a brief taxonomy of design characteristics and methods (Figure 4). A more comprehensive coverage of the field is provided by recent works and guidelines of Hoffman [210], Bartneck et al. [209], and Bethel et al. [211].

Three major study paradigms can be differentiated [212]: *Insight-driven studies* are directed towards developing general ideas or new theories in a problem context. Examples from the field of industrial cobot use are e.g. understanding (un-)desired job attributes to inform cobot deployment by interviewing assembly line workers [213], finding aspects relevant to trust in robots [50], or identifying challenges associated with cobot deployment [214], [215]. In contrast to the exploratory nature of insights-driven studies, a *design study* is specifically directed towards designing concrete robots or robot behaviours, e.g. during participatory work on demonstrators [204]. Finally, *hypothesis-driven studies* seek to objectively test hypotheses with statistically significant, numerical data. In the context of industrial applications, we can e.g. hypothesize that a mixed human-robot team will enhance productivity, job quality etc., compared to the baseline condition of the same process when performed by a human.

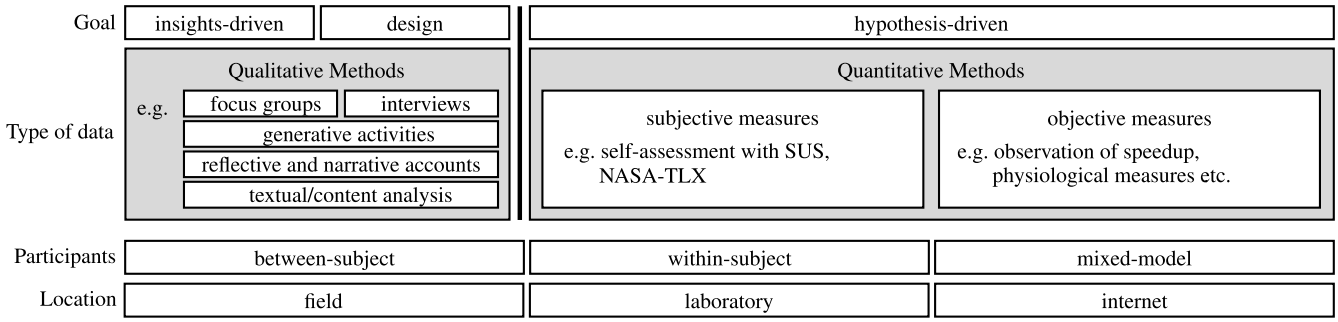

FIGURE 4. A brief taxonomy of goals and associated human-participant study design characteristics.

The above paradigms are pursued with qualitative and quantitative experimental procedures according to the type of data they provide [210], [212]. *Qualitative methods*, such as (semi-)structured interviews, focus groups, generative activities, or reflective and narrative accounts, provide textual participant responses, field notes, audio or video recordings to be interpreted by the researcher. The resulting data can typically not be expressed numerically [209] – qualitative methods are, hence, predominantly used for insights-driven and design studies [212]. By contrast, *quantitative methods* condense complex aspects into directly and objectively comparable numerical measures as needed for statistical hypothesis testing. The chosen measure defines the experimental procedure during quantitative studies: *Subjective measures* are gathered from participants by asking them to self-report their experience with the system, e.g. by answering questionnaires. By contrast, *objective measures* can be obtained through independent observations, e.g. measuring task durations or human physiological variables [216]. With most of the metrics discussed in this paper being numerical, Section III is a catalogue of subjective and objective quantitative measures.

Recruitment and involvement of participants are of crucial importance in human-participant studies. It is not only important to ensure a sufficient number of participants to achieve an appropriate level of significance (e.g. using an a priori power analysis [194]) but also how these participants are divided into groups to test a system under different conditions. With regard to this design dimension, a separation into between-subject, within-subject, and mixed-model studies is common [194]: In a *between-subjects study*, the participants are randomly divided into several groups. Each group will then take part in one variation of the experiment, and results will be compared across the groups (e.g. [125], [217], [218], [219], [220], [221]). By contrast, each subject experiences several experimental conditions in a randomized ordering in *within-subject* studies (e.g. [119], [197], [222], [223], [224], [225], [226], [227], [228]). Compared to between-subject designs, this approach enables comparisons between participants and allows the collection of more data per participant – this may, on the other hand, lead to habituation and fatigue effects [194]. Finally, *mixed-model factorial designs* (e.g. [229], [230]) combine the aforementioned designs by running a within-subject study (e.g. regarding interaction experience [229]) with each member of groups in a between-subjects study (e.g. regarding robot anthropomorphism [229]).

Independently of the assignment of participants to conditions, user studies can further be distinguished regarding the *location* where they are carried out: Experiments in the *field* with demonstrators of productive use-cases are still rather rare, especially when it comes to true task sharing rather than mere human-robot coexistence [205], [231], [232]. Respectively, studies in *laboratory* settings are predominant (e.g. [119], [197], [217], [218], [220], [221], [222], [223], [225], [228], [229], [230]) albeit partly seeking to replicate realistic industrial settings rather than relying on synthetic tasks (e.g. [219], [224], [233]). Laboratory studies can be conducted with physical or virtual prototypes. In the latter case, different parts of the robot system under test can be virtualized utilizing simulation: Virtual hardware can be accessed by participants to share tasks with robots using virtual reality techniques [197] (cf. Section IV-C). Likewise, the existence of software components can be simulated by substituting autonomous teaming capabilities with a researcher teleoperating an embodied robot in *Wizard-of-Oz* studies [220]. Finally, the potential for large, representative samples through crowdsourcing has motivated using the *internet* as a study context, e.g. by presenting videos [234] or by enabling the aforementioned interaction in virtual spaces [235], [236].

C. VIRTUAL COMMISSIONING

The term *virtual commissioning* has traditionally been used for procedures during which physical hardware is connected to a realtime-capable simulation system, e.g. for programming and testing Programmable Logic Controllers decoupled from the physical production system [237]. Lechler et al. have only recently pointed out that this technique also offers a high potential for applications beyond this early use-case, particularly for HRI [238] – indeed, simulations of robot as well as human behaviour have been part of several HRI experiments in recent years. This section frames corresponding

evaluation strategies in the context of virtual commissioning terminology. As we are focussing on settings in which humans and robots *actively* share a joint task (Section I), approaches where humans *passively* observe simulated robot behaviour (e.g. [227], [234]) will not be considered in depth. Following the nomenclature of Erős et al. we distinguish between *immersed* and *virtual* human-in-the-loop virtual commissioning [239].

1) IMMERSED HUMAN-IN-THE-LOOP VIRTUAL COMMISSIONING

Immersed Human-in-the-Loop virtual commissioning puts study participants into a virtual space where they can interact with a simulated robot system [239]. Humans perceive the virtual cobot through virtual reality (VR) hardware. We can here distinguish head-mounted displays (HMDs) [115], [198], [201], [240], [241], [242], [243] from less frequently used systems that project scenes onto walls of the room [196], [197], [244]. Participants' experience of the virtual world is not limited to the visual modality but can also be enriched with auditory cues, e.g.ambient noise in industrial facilities (e.g. [240], [243]). The control loop is closed by providing humans with an input channel to interact with the virtual workspace. This can be achieved with the controllers, which are usually deployed with VR HMDs (e.g. [198], [201], [240], [242]), with external cameras (e.g. [243]), or with marker-based tracking systems (e.g. [115], [197], [245]). More precise information, particularly on human hand motions, is provided by data gloves [245], or by attaching a hand tracking sensor to the HMD [241]. These input devices enable a variety of collaborative jobs, ranging from pick-and-place tasks with object handovers between agents [115], [197], [242], [246] to high-fidelity assembly settings [198], [240], [241], [243]. Virtual environments may also allow humans to move on larger shop floors (e.g. [242], [246]). In addition to task sharing, even kinaesthetic robot teaching can be virtually tested using haptic devices with force feedback [244].

Human interaction influences the virtual environment, and the control and planning algorithms under test produce corresponding robot reactions – just as in human-subject studies with physical robot prototypes. Accordingly, the design dimensions outlined in Section IV-B equally apply to user studies conducted by Immersed Human-In-The-Loop Virtual Commissioning. Corresponding studies mostly seek to gather quantitative data (e.g. [53], [115], [196], [197], [198], [240], [241], [242], [243], [247]), but virtual reality also enables insights-driven research [248]. A major difference is that human safety during experiments is not an issue in virtual environments as opposed to physical interaction in laboratory setups. Consequently, we found that (possibly dangerous) variations of robot trajectories in terms of speed or unexpected, jerky movements are a frequently manipulated variable [115], [196], [197], [198], [242], either directly or indirectly when testing different cobot safety mechanisms [243]. Other variables are, e.g., the use of different communication channels [240], [247], robot morphology [197], or unforeseen events in the surrounding of the shared task [246]. Dependent measures cover the full spectrum of metrics outlined in Section III: Productivity in terms of completion times, errors, and fluency have been observed [115], [196], [198], [240], [243] as well as the flexibility of humans regarding changes to the robot working speed [198]. In terms of job quality, physical ergonomics scores have been calculated from skeletal tracking data [241], and cognitive workload related to stress has been evaluated with corresponding questionnaires (e.g. [115], [196], [247]). Lastly, the safety of planned robot motions can be tested against true human motion data (e.g. [201], [240], [243]). Constructs related to psychological safety, including anxiety, trust, and behavioural patterns (e.g. leaning back or stepping away from the robot), have also been observed in VR-based experiments [53], [196], [197], [198], [242]. All aforementioned studies have been conducted in laboratories. However, VR technology is also a suitable medium for crowdsourcing and remote participation in HRI experiments via the internet, i.e. without having to invite subjects to the laboratory [235].

2) VIRTUAL HUMAN-IN-THE-LOOP VIRTUAL COMMISSIONING

Contrasting to Immersed Human-in-the-Loop virtual commissioning, *Virtual Human-in-the-Loop Virtual Commissioning* relies not only on virtual robot hardware but also on simulated human behaviour – human subjects are thus not part of corresponding experiments [239], but the HRI is fully simulated. Related approaches can be classified according to the expressiveness of underlying *digital human models* (DHMs) used to replicate human behaviour:

Most DHMs emphasize *physical* aspects of human action [249], e.g. regarding motion times and ergonomics. Accordingly, systems for simulating manual work (e.g. ema work designer [250]; IPS IMMA[1]) have been used to analyse cycle times and ergonomic metrics of collaborative workflows [4], [203], [251]. In contrast to these 3D simulations with realistically animated manikins and robots, 2D simulations of workers and mobile robots on the shop floor [252] or of hand motions above a working surface plane [242], [253], [254] have been proposed to estimate task times. In addition to productivity- and ergonomics-related metrics, simulations are predestined to investigate safety mechanisms without actually endangering human subjects. To this end, expected contact forces can be estimated with biomechanical collision models (e.g. according to ISO/TS 15066 [36]) to design workstations that are prepared to pass the risk assessment afterwards [203], [255]. Particularly approaches based on commercial simulation tools are designed to iteratively draft and test candidate HRC workflows by precisely entering work items for all involved agents. In contrast, Antakli et al. [202] have proposed a simulation architecture for more interactive testing with coupled agent behaviours: Production planners

[1] https://industrialpathsolutions.se (Date accessed: 2022/03/28)

can here create different situations by manipulating objects and agent states at simulation runtime, hence influencing the course of actions emerging from a near-optimal optimization scheme and human motion synthesis.

Cognitive aspects of human behaviour are hard to model and have less frequently been addressed in the context of industrial HRC evaluation. There are practical models capturing individual factors and dependencies, e.g. between

- fatigue, learning, and human error rates [256]
- robot performance, task complexity, human physical and cognitive workload [257]
- trust and human-robot teaming performance [258]
- eye gaze and hand-reaching motions [259]
- consecutive decisions when choosing from several assembly steps yet to be done [260].

These can be used as components in models of high-level human decision behaviour to simulate non-deterministic but plausible human action in industrial settings [10], [261]. Recent models include behaviours such as inter-agent communication and leaving the workstation for a break [261], e.g. by relating transition probabilities in a Markov Decision Process to human fatigue [261], [262] or frustration [261] accumulated over time.

D. ANALYTICAL MODELS

The final evaluation strategy that we are going to survey makes use of analytical models. Such models seek to encode parametric relationships in sets of mathematical equations. They are primarily applied to observe cobot impacts on productivity metrics when varying different economic parameters. In this class, process-independent and process-oriented approaches are distinguishable. *Process-independent models* are designed to estimate the minimum productivity gain (e.g. in terms of assembly line throughput [263]) that must be achieved to justify the introduction of cobots [263], [264]. To this end, expected cycle time reductions resulting from teamwork are set in relation to the costs of cobot use. Beyond initial hardware acquisition expenditures, the cost model proposed by Calvo and Gil [264] also reflects product changes during the operative lifetime, wage raises over time, and social costs, such as welfare support for human workers replaced by robots. In contrast, the approach of Cohen et al. puts a stronger emphasis on cobot-based compensation of productivity losses emerging from the temporal absenteeism of experienced workers and replacement with less experienced ones in assembly lines [263].

The applicability of such high-level models presumes that estimates of cycle times for manual processes and particularly of human-robot teams are already available. Although this question is strongly product- and process-dependent, several authors have shown how analytical *process-oriented models* for hybrid workstations can be shaped. On this level of system analysis, deterministic and stochastic models have been proposed. *Deterministic models* such as the one proposed by Faccio et al. [265] are applicable to a class of processes in which individual process steps do not depend on each other and are assumed to have equal durations. More generally, arbitrary processes with precedence relations between operations can be analysed by modelling human-robot teamwork as a multi-agent scheduling problem [266]. This model can then be solved for different HRC settings (e.g. varying number of humans and robots, discrepancy between human and robot working speed, percentage of process steps that the robot is capable of etc.), yielding the optimal workflow and time savings compared to manual work. A consequence of optimality is that this model does not capture the variance of dynamic, flexible teams across different runs of the same process. In contrast, *stochastic models* have been proposed to account for uncertainties in manual operations. Similar to the aforementioned model of Faccio et al., these models rely on a limited process pattern. But they assume process step durations which follow exponential distributions [54] or gamma distributions [267] – it is this way possible to calculate the expected value of the overall process duration and even derive formulas for the probability of a product to be finished within given time bounds [267], [268].

TABLE 3. Classification of evaluation strategies regarding their ease of use, versatility, and impact on experiment reproducibility (👎= low, 👍 = medium, 👍= high, – = not applicable).

	Ease of Use	Reproducibility	Versatility	Participant Well-Being
Research Demonstrators	👍	👎	👎	–
Human-Participant Studies	👎	👎	👍	👍
Immersed-HITL VC	👎	👎	👍	👍
Virtual-HITL VC	👍	👍	👍	–
Analytical Models	👍	👍	👎	–

E. DISCUSSION

From our point of view, which is based on prior considerations on benchmarking in the computing domain [269], there are three major goals to be satisfied by benchmark protocols to become an established part of the scientific method:

- *Ease of use*: Since benchmarks are merely a tool to gather the data needed for investigating research questions, they should be designed as researcher-friendly, easy-to-use testbeds which enable cost-effective, time-efficient, and scalable experiments.
- *Reproducibility*: It must be feasible to repeat prior experiments to the greatest possible extent, as this renders research results transparent.
- *Versatility*: Benchmarking protocols should be designed to gain information flexibly regarding various constructs and relevant metrics to ensure versatility and foster high acceptance in the community.

TABLE 4. Suitability of different evaluation strategies for raising metrics as categorized in Section III.

			Flexibility		Job Quality		Safety	
	Feasibility	Productivity	Team	Task	physical	cognitive	physical	psychol.
Research Demonstrators	●	○	○	○	○	○	○	○
Human-Participant Studies	●	◑	◑	●	●	●	○	●
Immersed-HITL VC	◑	◑	◑	◑	●	●	●	●
Virtual-HITL VC	○	●	●	○	●	◑	●	○
Analytical Models	○	●	○	○	○	○	○	○

Beyond this general, practicality-oriented view, another important aspect of experimental setups must be discussed in the particular context of humans interacting with robots:

- *Participant Well-Being*: During any benchmark or experimental procedure, it must be easy to ensure that human participants are safe and not harmed at any time.

Each of the strategies in Sections IV-A to IV-D has individual properties, (dis)advantages, and best practices which influence the above aspects – corresponding dependencies are discussed in the below sections. A brief summary of the resulting classification is given in Table 3, with the achievable versatility in terms of metrics covered by individual strategies being further elaborated in Table 4.

1) EASE OF USE

Building *research demonstrators* usually goes along with significant engineering efforts, thus leading to limitations in the ease of use: Intelligent robots require complex software stacks to expose all skills necessary to collaborate with humans. When seeking to evaluate novel planning and interaction methods, it is e.g. also necessary to implement state-of-the-art vision and manipulation algorithms – as these system components are a necessity which mostly does not directly contribute to individual research goals, corresponding software and hardware are often closed-source, simplified, or specifically tailored to work in individual laboratories. This does not only negatively impact reproducibility but also leads to rather heavy-to-use systems [14].

When conducting *human-participant studies* with (physical) cobot setups, challenges beyond engineering efforts and unstable, error-prone prototypes further reduce the *ease of use*: (i) Particularly when relying on participant self-assessment with questionnaires, various influences on human behaviour with potential impacts on the validity of results must be considered (e.g. social desirability bias; novelty vs. habituation effects; side effects outside the study protocol, such as robot failure [14], [194]). Accordingly, conducting high-quality human-participant studies is challenging, and there has lately been profound criticism of a lack of methodological rigour in the field. This concerns a lack of reproducibility, critical conclusions from too small populations [14], a strong focus on convenience samples [210] with corresponding biases (e.g. regarding participants' age [270]), or incorrect design and statistical testing of Likert scales and associated data [271]. (ii) With field demonstrators for human-robot task sharing still being rare [231], [232], most experiments take place in laboratories [270] and are often conducted with synthetic model sets and tasks – this raises further questions regarding the trade-off between experimental control and external/ecological validity (see Section V). All in all, achieving rigour in the design, execution, and reporting of human-participant studies is a complex task. We therefore want to refer the reader to further literature which introduces best practices in depth [209], [210], [211], [271], [272].

Immersed-HITL Virtual Commissioning is a special case of human participant studies with virtual rather than physical robot prototype systems. Certainly, not needing to build possibly expensive physical prototypes and not having to ensure the safety of participants compared to the laboratory operation of physical robots are favourable aspects regarding the *ease of use*. Despite these advantages, we still judge Immersed-HITL VC to be a similarly complex, hard-to-use use evaluation strategy as studies with physical robots: beyond the issues for rigorous studies, experiments conducted in VR raise the question of transferability of results from the virtual to the 'real' world. Although VR experiments are already frequently used, this assumption is still discussed in the literature (e.g. by Wijnen et al. [273]). Transferability should thus not generally be assumed. It is widely accepted that *presence*, i.e. the feeling or illusion of actually being in an immersive virtual environment [274], is a key prerequisite to ensure realistic participant responses when subjects interact with a virtual robot [275]. Presence is therefore often measured and discussed as a part of VR user studies to justify the validity of results (e.g. [53], [115], [196], [198], [246], [247], [276]). A comprehensive list of available questionnaires covering the presence concept has been compiled by Schwind et al. [277]. With presence being related to the fidelity [274], [278] and validity [278] of a virtual world, these aspects should already be considered during a structured design phase for high-quality virtual environments [245].

The complexities associated with human subject handling do not arise for *Virtual-HITL Virtual Commissioning* with simulated humans: Building and running fully

virtual human-robot tasks is supported by commercial tooling (see [279] for a listing) as well as by freely available simulation platforms with physics, sensor simulation, and human motion animation capabilities (e.g. gazebo,[2] Webots[3]). Once a simulation environment has been prepared, adjustments to the layout, to the human-robot task assignment etc. can be arbitrarily evaluated. It has been shown that workflow variations can also be generated automatically [10], [261], hence enabling large-scale experiments without needing to recruit human subjects and even without an experimenter's active supervision. These factors increase the ease of use compared to human-participant studies with physical or virtual prototypes. However, we still consider the gathering and integration of realistic simulation environment assets (CAD models) a task which needs a certain degree of expert knowledge and experience (e.g. when modelling complex assembly processes with professional tools as the ema Work Designer [250]). This sets a limit to the ease of use.

Lastly, we consider *Analytical Models* – the most experimenter-friendly tooling in this survey. Such models can be compiled by 'pen-and-paper'-work without realizing complex robot software stacks. They provide reliable and verifiable results without effortful system implementations or demanding user studies. Different scenarios can easily be evaluated within the bounds of aspects considered in the model by determining suitable input parameter values and solving for the output metrics.

2) REPRODUCIBILITY

The aforementioned aspects related to closed-source, simplified, and specifically tailored hardware and software of laboratory *research demonstrators* render reproduction of most robotics experiments hardly feasible [280] – they are seen as a major source of the so-called 'replication crisis' in HRI [14]. We consequently consider the overall reproducibility of research demonstrators as low at the time being (Table 3). These issues directly propagate to *human-participant studies* as these are usually based on prototype cobot implementations. Yet there is even more to reproducible human-participant studies than the reproduction of the mere experimental platform (hardware and software): It is here moreover necessary to make the experimental setup (e.g. the benchmark task, cf. Section V) as well as the experimental procedure (study design, questionnaires etc.) available [281]. These issues are addressed by initiatives to foster publications with extended, detailed information on the used hardware and software implementation: So-called 'R-Articles' must be accompanied by mandatory, in-depth system descriptions, code, and further data relevant for reproducing experimental setups [282]. This step towards more transparent experiments is supported by online platforms to publish the required data (e.g. CodeOcean,[4] IEEE DataPort[5]). From a technical point of view, the situation may be enhanced in the future by unified architectures for experimental cobot systems [283] and by the use of containerization techniques [284], [285]. Yet, to the best of our knowledge, these approaches have not yet been applied to complex HRI user studies – we hence classify the reproducibility of human-participant studies as low.

When transitioning from human-participant studies with physical prototypes to *Immersed-HITL VC*, the experimental platform is reduced to software to be built for and run on commercially available standard VR hardware. Beyond this reduction of required hardware, questionnaires (see [286] for design guidelines of in-VR-questionnaires) and the overall experimental procedure can be embedded into the code base (e.g. by means of 'mini games' [287]). In our ranking, the reproducibility that such integrated VR experiments could offer is only superseded by fully simulated *Virtual-HITL VC* and *analytical models*. Structured containerization of simulation components has been proposed [288], and closed-form analytical performance models may not even need additional materials for reproduction. In both cases, it may not only be possible to easily reproduce the experiment but even to precisely replicate and verify prior results.

3) VERSATILITY

The experimental versatility enabled by building a *research demonstrator* is limited to proving technical feasibility of a novel approach. Demonstrators can certainly be used to measure technical aspects of individual system components (e.g. object classification accuracy, positional accuracy etc.) – but to acquire any further metrics, laboratory prototypes must be embedded into *human-participant studies* to measure aspects related to HRI. The user studies surveyed in Section IV-B report metrics across all relevant categories outlined (Table 4). The versatility when conducting human-participant studies is, thus, very high, and metrics to be raised are only bounded by two limiting factors: (i) The physical integrity of participants is a requirement during human-participant studies, and physical safety metrics can thus not be evaluated. (ii) Guiding participants through a laboratory HRI user study usually needs time to introduce the topic, to perform tasks together with a cobot, and to query feedback with questionnaires. This expenditure of time is limited by the capacities of experimenters and participants. In consequence, the number of different workflows and scenarios to be tested is equally limited. This restricts the applicability of human-participant studies to productivity metrics under the influence of team flexibility.

These limitations can be overcome with Virtual Commissioning methods. When applying *Immersed-HITL VC*, user studies can measure physical safety metrics (e.g. the separation distance) without taking the risk of harmful human-robot collisions. VR moreover enables crowdsourcing, which

[2]https://gazebosim.org/home (Date accessed: 2022/06/15)
[3]https://cyberbotics.com/ (Date accessed: 2022/06/15)
[4]https://www.codeocean.com (Date accessed: 2022/05/16)
[5]https://ieee-dataport.org (Date accessed: 2022/05/16)

can help to increase sample sizes and reduce the required time by moving user studies to participants' homes [235]. *Virtual-HITL VC* enables experiments ranging from tests of single, fixed workflows [203], [251] to large-scale evaluations of automatically varied human-robot co-working processes [10], [261] and structured situation coverage [289]; not building experiments around human subjects as a limited resource means in particular that team flexibility and the effects of dynamic task sharing can fully be covered. Yet modelling plausible human behaviour which reflects the indeterminism in human actions is still an open challenge [290]. As a consequence, we see the simulation-based evaluation of qualitative human factors (cognitive ergonomics, psychological safety etc.) still in its infancy. Due to the high importance of human factors metrics, we assign a medium versatility to this approach despite the broad range of scenarios and metrics covered. Still, Virtual-HITL VC is already a versatile tool, which should especially be considered during early preparatory steps of enabling technologies research prior to future detailed human subject experiments [291].

Compared to the above strategies, *Analytical Models* are less versatile. To our knowledge, existing models target estimating productivity gains as the key output metric. Aspects related to team flexibility (e.g. absenteeism of workers [263], number of humans/robots in the team [253], [266]) or to task flexibility (e.g. switching costs for programming and exchanging hardware components [264], structure of the task [254]) are not evaluated but used as input parameters or constants tailored to specific use-cases. The high-level view on whole processes or abstract process steps enables complex economic considerations on the long-term implications of human-robot teaming [264] – yet, concrete system details can not be modelled on this high level of abstraction, hence preventing detailed evaluations of human factors or safety.

4) PARTICIPANT WELL-BEING

The issue of participants' well-being emerges for the two human-subject-based strategies. It is here important to guarantee that human subjects are not harmed, neither psychologically nor physically, for ethical and legal reasons. When working with hardware prototypes, physical integrity can theoretically be retained by safety mechanisms in line with relevant norms and standards on occupational safety (see e.g. [60] for a comprehensive overview). Yet compliance with the still rigid risk assessment procedures is hard to realize for flexible robot systems which plan and behave dynamically [292]. This leaves us with the strategy of an experimenter carefully overseeing the situation and operating the dead man's button [228]. When combined with lightweight, intrinsically less harmful robots as practised in recent user studies (e.g. [207], [218], [219], [222], [223], [225], [229]), this strategy is an acceptable solution. Simulator sickness is a problem related to Immersed-Human-in-the-Loop VC without a similarly common solution. Despite the technical improvements to VR hardware in recent years, this phenomenon still frequently affects humans during and even after using VR hardware [293]. The influence of technical and temporal aspects [293] as well as of the content displayed in VR [294] should therefore be considered when designing the virtual HRI environment. Established tools such as the Simulator Sickness Questionnaire (SSQ) [295] can be used to validate one's setup with regard to this aspect. Additionally, a debriefing phase with each participant can help to identify unforeseen harms and offer assistance if needed [210]. Overall, any experimental procedure (virtual or with physical robots) should be reviewed by institutional ethics committees or review boards (see e.g. [211]). This is not only a necessary prerequisite for publishing results with some venues, but it will also ensure a holistic view on potential dangers that researchers used to robots might overlook.

V. BENCHMARK TASKS

With metrics (Section III) and experimental strategies (Section IV), we have so far covered two main components of benchmarks for human-cobot teamwork. The definition of the actual task is the last component. More formally, *benchmark problems* in terms of object model sets and associated actions are "designed or used to establish a point of comparison for the performance or effectiveness of something" [296] are needed. Accordingly, several model sets have already been proposed in the general field of robotics research. They range from household settings [16], [17], [297] to industrial bin-picking [298], [299] and assembly scenarios [15], [300], [301]. Corresponding tasks are mainly designed to challenge robot grasping and manipulation skills based on robot performance metrics (e.g. time to completion and success rates [15]). Raising such metrics is also important in the field of HRI (Section III-A). Yet, the scalability of manipulation complexity is not sufficient to evoke human- and team-related effects with an impact on job quality, safety etc. By contrast, appropriate reference tasks for collaborative scenarios require scalability regarding (i) individual agent capabilities and contributions to the tasks, (ii) complexity of the required interaction and coordination, also in terms of communication, and (iii) applicability to the different teaming modes still under investigation (e.g. predefined task allocation, negotiation, implicit mutual adaptation) [302]. There have been a few attempts to define reference collaboration tasks and model sets with these requirements in mind (Section V-A). Beyond these works, we have comprehensively surveyed the individual tasks used in previous experiments to provide further inspiration (Section V-B).

A. DEDICATED MODEL SETS AND REFERENCE TASKS

The number of publications with an explicit focus on dedicated model sets and reference tasks for human-robot task-sharing benchmarks is rather limited. Zeylikman et al.have proposed a modular model set including plywood panels, dowels, and freely available 3D-printed connectors [302]. From these components, differently complex pieces of

TABLE 5. Tasks and objects used in joint action experiments, grouped by domain and experiment location (in physical space or virtual reality).

	physical	virtual
Pick-and-Place		
• geometric primitives	[119] [217] [25] [218] [9] [303] [223] [10] [125]	[246] [242] [115] [197]
• other objects	[70] [218] [304]	
Toy Assembly		
• interlocking plastic bricks	[7] [229] [90] [110] [226] [305]	
• other toys	[174] [88] [306]	
Product Mock-up Assembly		
• furniture assembly	[307] [11] [302] [308] [309]	
• gear meshing/gear boxes	[25], [72], [219]	
• electrical circuitry	[25], [310]	
• other artificial products	[222] [311] [312] [313] [314] [12] [315]	
Realistic Product Assembly		
• placing/fastening screws	[65], [221], [316]	[198]
• (partial) engine assembly	[233], [317]	
• other realistic processes	[318] [8] [319] [320] [176] [224]	[240] [241] [243]

furniture can be assembled. Besides this scalability of task complexity and duration, a mixture of actions feasible for both human and robot agents (e.g. bringing and holding certain parts) and actions which exclusively require human dexterity (e.g. screwing) enable scaling the interaction in terms of role assignment. Our prior work [25] has a similar focus on easy-to-reproduce, task-centred interaction. The model set spans a broader range of domains (simple building blocks, abstracted electrical circuitry, gear meshing). All parts are designed for robust manipulation by humans and particularly by robots using specific shapes and adhesive forces. This way, confounds as a consequence of robot failure during user studies are actively prevented. Contrasting to these assembly-inspired model sets [25], [302], the task described by Sarthou et al.is based on the 'Director Task' as known from psychology studies and hence fosters cognitive and behavioural aspects more strongly based [303]. Involved agents are here facing each other with a shelf in between. From this shelf, cube-shaped objects have to be picked – this approach offers less scalability regarding task-centred complexity (e.g. coordination due to assembly precedence relations [25]), but intrinsically challenges referential communication, perspective taking etc.

B. TASKS USED IN JOINT ACTION EXPERIMENTS

In addition to the initial attempts towards establishing reference tasks as common ground for comparable, repeatable experiments (Section V-A), a broad range of tasks has been used in HRI experiments. We have clustered those tasks by domain in Table 5 which fall within the scope of this survey, i.e. which target scenarios in which subtasks are allocated to different agents in line with our definitions in Section I:

Tasks during which agents have to *pick-and-place* several objects from start to goal locations are frequently used in experiments with physical robots and in virtual reality. The parts to manipulate range from primitively shaped, often distinctly coloured objects (e.g. [25], [125], [223], [303]) to everyday life objects (e.g. apples [304], USB keys [9]) and paper cut-outs [217]. Tasks composed of pick-and-place subtasks are often intended to represent packaging jobs as an underlying use-case [9], [10], [217], [304]. They may also incorporate parts stacking as a strongly abstracted form of assembly (e.g. [10], [25], [115]).

Another widespread category of tasks, which we identified, is *toy assemblies*. Building 2D (e.g. [7], [226], [305]) or 3D structures (e.g. [90]) from interlocking plastic bricks is similar to pick-and-place tasks but requires a slightly higher level of manipulation skills, especially on the robot side. Other construction toy sets (e.g.'Baufix' with wooden screws, nuts, and bolts [174], [306]) bring experiments closer to the next group, which we have named *product mock-up assemblies*. This category covers artificial products intended to simulate realistic assembly processes. In contrast to the real process, products are here built from specially constructed, often strongly simplified parts. Several HRI experiments have applied human-robot teamwork to scaled pieces of furniture [11], [302], [307], [308], [309]. Other exemplary tasks are gear meshing and gearboxes [72], [219], electrical circuitry [310], a jet engine mock-up [315], bicycle subassembly [314], sanding machine [313], or flange assembly from metal parts and standard screws [12]. In addition to these tasks created by abstracting systematically from real products and production processes, fully synthetic tasks such as the Cranfield Benchmark [312], Bourjault's Pen [311], or simple 3-component products [222] have been used.

Compared to product mock-up assemblies, our final category of *realistic product assemblies* summarizes tasks which involve joint work with real parts taken from industrial processes. Within this scope, we found a group of experimental setups in which humans and robots needed to coordinate while placing and fastening screws [65], [198], [221], [316]. Aside from this cluster, use-cases are highly individual. They include the assembly of desktop PCs [8], candy tins [224], emergency buttons [318], a filament winding head [241], car engine sub-assemblies [233], [317], pin-back buttons [240], USB adapters [320], or carbon fibre shells [243].

C. DISCUSSION

When investigating aspects of human-robot task sharing, the model set and tasks used are strongly linked with the chosen experimental strategy (Section IV). We will therefore discuss the matter of benchmark tasks in the context of the same goals as defined for experimental strategies (ease of use, reproducibility, and versatility; see Section IV-E):

Ease of Use: Research prototypes in the HRI field are often based on simplifications to ease robot system implementation (see subsection 'Realistic Product Assembly' in Table 5). *Robot vision* is often supported by attaching a fiducial marker

to each object (e.g. [302], [303], [311], [312]) or by using distinctly coloured objects for more robust segmentation (e.g. [7], [10], [25], [223]). Similarly, *robot manipulation* can be rendered more robust based on the design of parts and supporting structures in the environment: simply shaped parts geometries (e.g.cubes [223], [303]) or extra handles added to objects [25], [72], [218] facilitate obtaining a stable grasp. Markings on the workbench (e.g.grids [222], [223]), as well as fixtures, foster more precise parts placing and alignment [9], [72], [176], [219], [222], [307], [310]. Furthermore, embedding magnets into objects can help to reduce the required positioning precision and enhance the stability of jointly built structures [25], [125].

Certainly, all the aforementioned simplifications mean stepping away from realistic tasks and cobot use-cases – realism and fidelity are traded for robustness and, hence, increased experimental control. This abstraction step is an important feature, particularly in the context of user studies with physical prototypes: unintended robot failure due to unstable robot vision and manipulation would here mean confounds. These confounds reduce internal validity and can even render samples invalid. However, experiments with abstract model sets and tasks raise the question of validity and transferability of results to the real world. Regarding this question, we can draw on insights from experimenting with *synthetic task environments* in the field of human factors research: Synthetic tasks inherit the relevant functional relationships from real-world tasks [321], but they have reduced *physical fidelity* regarding the equipment, environment etc. Still, investigations related to high-level human skills (e.g. teamwork, dynamic problem-solving) can be valid as long as they were conducted with high *psychological fidelity* regarding functional, cognitive, and construct-related aspects of the tasks under consideration [322]. To this end, synthetic tasks should emerge from a systematic abstraction and validation process based on identifying research objectives and the concrete field of practice [323].

Reproducibility: Highly realistic research demonstrators are important to prove the value of human-robot teaming in practice. However, highly realistic assembly station environments (e.g. [319], [320]) are hard to replicate due to high costs and a lack of information and plans in corresponding publications. Even if experiments are not conducted in fully developed assembly cells, replication can be prevented if the parts involved in joint tasks are taken from industrial processes (e.g. [233], [317]) and, hence, are not broadly available. As discussed in Section IV-E, these problems can be reduced by shifting experiments into virtual spaces. In case of user studies with physical demonstrators, synthetic tasks offer further beneficial properties also with regard to reproducibility: Product mock-ups can easily be made with 3D printers [25], [72], [219], [302], [311], [312], [315] which are a broadly available resource in the meantime. Unfortunately, prior publications with interesting model sets are not always accompanied by the required CAD models (e.g. [72], [219], [315]) – the replication of experimental setups could, hence, be strengthened by using specifically designed, online-available model sets [25], [302] (Section V-A), or by referring to online repositories with free CAD models as e.g.done by Cramer et al. [311]. As an alternative to 3D printing, parts as standard aluminium profiles [313], [314] or toy sets [7], [90], [226], [305], [306] are commercially available and easy to acquire. When relying on such parts, it is important from a reproducibility point of view to provide precise measures of screws and other metal components (as e.g.in [12]), to report on the concrete workspace layout (e.g.in [7], [217], [304], [305]), and to specify the patterns or structures to assemble (e.g. [7], [90], [223], [226], [305], [306]) with according referential graphics or photographs.

Versatility: Lastly, we will discuss properties which tasks should possess to be suited as a versatile-to-use benchmark problem for different HRI experiments. Versatility is achieved if a benchmark problem has variables which can be manipulated to create individual tasks with differently scaled characteristics, ideally related to various constructs and metrics under investigation. We extracted the following important dimensions of scalability from the tasks we considered in our analysis (Table 5): It is important to be able to scale the *amount of work* of a task. This can easily be achieved in the pick-and-place or toy assembly domains – additional parts and associated subtasks can here be added as needed. This also opens the possibility of arranging parts in the workspace and creating settings that encourage parallel working or provoke conflicts during reaching motions in spatially narrow situations [10], [223]. Similarly, product mock-ups with a higher degree of abstraction and a design specifically focused on scalability enable various options to configure tasks (e.g. [25], [302]). Other mock-ups (e.g. [72], [219]), as well as realistic demonstrators and case studies with high fidelity (e.g. [319], [320]), replicate exactly one product and can thus hardly be scaled regarding the necessary subtasks. A necessity to employ individual agents' *capabilities* can be created implicitly by adding particularly heavy parts (e.g. [9], [233], [319]). This also establishes a link to embedding physical contact between humans and robots by means of hand-guiding [233], [319]. Another way of implicitly introducing complementary capabilities is limiting robot manipulation skills to certain parts [9], [72], [233], [302], [310], [313]. Alternatively, capability-aligned contributions within an assembly process can explicitly be enforced by assigning a specific sort of process steps to certain agents (e.g.one teammate placing and the other fastening screws [65], [198], [221], [316]). It is here common to restrict robots to picking/placing/handing over parts [12], [174], [240], [306], [307], [308], [315], [317], [324] and holding sub-assemblies, whereas the human partner performs dexterous manipulations [241], [243], [309], [314], [324] or operates tools [240], [317]. On the one hand, embedding this distribution of work statically into experimental tasks has two advantages: it reflects the situation 'as is' with robots still

being less skilled than humans in dexterous manipulation; and it contributes to the ease of use as picking, placing, and holding actions can easily be implemented and enable robust execution during user studies. On the other hand, fixed roles can limit the *interaction* to a single, fixed task allocation [176], [219], [224], [319], [320] or even mean that robots merely deliver kits for humans to assemble without further interaction during the actual assembly process (e.g. [88], [110], [229]). As opposed to these limitations, it has been shown that especially the easier-to-implement pick-and-place tasks [125], [217], [218], [223], [303], [304], toy assemblies [7], [226], [305], and abstract product mock-ups [11], [25] can also be realized in a way that enables the other extreme case of equal, symmetric capabilities. In consequence, these tasks can be used to benchmark the full range of HRC modes from statically planned optimal schedules to fixed 'leader-follower' role distributions and, ultimately, fully dynamic mutual adaption among equal peers. Independently of the HRC mode under investigation, the cognitive load can be scaled by superimposing an additional cognitive task on the original assembly task (e.g.solving Towers of Hanoi [222], product quality inspection [196]).

VI. CONCLUSION

Standardized benchmarks are an important foundation of reproducible research and comparable performance evaluations. From our point of view, such benchmarks are yet to emerge in the field of collaborative human-robot task sharing. Towards this target, our survey seeks to give an overview of aspects related to HRI experiments. Compared to prior literature reviews with a focus on HRI metrics, we provide a broader overview of the field: we have surveyed metrics more specifically in the context of the currently most frequent use-case of human-robot collaboration in industrial settings by investigating their suitability to measure productivity, flexibility, job quality, and safety in Section III. Evaluation strategies to raise these metrics, particularly when dynamic teaming approaches and mutual adaptation would require large numbers of test runs or subjects, are discussed in Section IV (e.g.human-participant studies, variations of virtual commissioning). Lastly, we have gathered dedicated object model sets and tasks previously used in HRI experiments in Section V. We hope this comprehensive overview will serve the community as a starting point and inspiration for the future design of scalable benchmark problems, protocols, and evaluation procedures.

REFERENCES

[1] T. Kopp, M. Baumgartner, and S. Kinkel, "Success factors for introducing industrial human–robot interaction in practice: An empirically driven framework," *Int. J. Adv. Manuf. Technol.*, vol. 112, nos. 3–4, pp. 685–704, Jan. 2021.

[2] G. Evangelou, N. Dimitropoulos, G. Michalos, and S. Makris, "An approach for task and action planning in human–robot collaborative cells using AI," *Proc. CIRP*, vol. 97, pp. 476–481, Jan. 2021.

[3] Y. Y. Liau and K. Ryu, "Task allocation in human–robot collaboration (HRC) based on task characteristics and agent capability for mold assembly," *Proc. Manuf.*, vol. 51, pp. 179–186, May 2020.

[4] V. Weßkamp, T. Seckelmann, A. Barthelmey, M. Kaiser, K. Lemmerz, P. Glogowski, B. Kuhlenkötter, and J. Deuse, "Development of a sociotechnical planning system for human–robot interaction in assembly systems focusing on small and medium-sized enterprises," *Proc. CIRP*, vol. 81, pp. 1284–1289, Jan. 2019.

[5] M. Pearce, B. Mutlu, J. Shah, and R. Radwin, "Optimizing makespan and ergonomics in integrating collaborative robots into manufacturing processes," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 4, pp. 1772–1784, Oct. 2018.

[6] L. Johannsmeier and S. Haddadin, "A hierarchical human–robot interaction-planning framework for task allocation in collaborative industrial assembly processes," *IEEE Robot. Autom. Lett.*, vol. 2, no. 1, pp. 41–48, Jan. 2017.

[7] B. Hmedan, D. Kilgus, H. Fiorino, A. Landry, and D. Pellier, "Adapting Cobot behavior to human task ordering variability for assembly tasks," in *Proc. Int. FLAIRS Conf.*, vol. 35, 2022, pp. 1–6.

[8] Y. Cheng, L. Sun, and M. Tomizuka, "Human-aware robot task planning based on a hierarchical task model," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1136–1143, Apr. 2021.

[9] A. Pupa, W. Van Dijk, and C. Secchi, "A human-centered dynamic scheduling architecture for collaborative application," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4736–4743, Jul. 2021.

[10] D. Riedelbauch, "Dynamic task sharing for flexible human–robot teaming under partial workspace observability," Ph.D. dissertation, Dept. Comput. Sci., Univ. Bayreuth, Bayreuth, Germany, 2020.

[11] K. Darvish, E. Simetti, F. Mastrogiovanni, and G. Casalino, "A hierarchical architecture for human–robot cooperation processes," *IEEE Trans. Robot.*, vol. 37, no. 2, pp. 567–586, Apr. 2021.

[12] G. Bruno and D. Antonelli, "Dynamic task classification and assignment for the management of human–robot collaborative teams in workcells," *Int. J. Adv. Manuf. Technol.*, vol. 98, nos. 9–12, pp. 2415–2427, Oct. 2018.

[13] N. Nikolakis, K. Sipsas, P. Tsarouchi, and S. Makris, "On a shared human–robot task scheduling and online re-scheduling," *Proc. CIRP*, vol. 78, pp. 237–242, Sep. 2018.

[14] K. Belhassein, G. Buisan, A. Clodic, and R. Alami, "Towards methodological principles for user studies in human–robot interaction," in *Proc. Test Methods Metrics Effective HRI Collaborative Hum.-Robot Teams Workshop, ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2019, pp. 1–7.

[15] K. Kimble, K. Van Wyk, J. Falco, E. Messina, Y. Sun, M. Shibata, W. Uemura, and Y. Yokokohji, "Benchmarking protocols for evaluating small parts robotic assembly systems," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 883–889, Apr. 2020.

[16] J. Leitner, A. W. Tow, N. Sunderhauf, J. E. Dean, J. W. Durham, M. Cooper, M. Eich, C. Lehnert, R. Mangels, C. McCool, P. T. Kujala, L. Nicholson, T. Pham, J. Sergeant, L. Wu, F. Zhang, B. Upcroft, and P. Corke, "The ACRV picking benchmark: A robotic shelf picking benchmark to foster reproducible research," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 4705–4712.

[17] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set," *IEEE Robot. Autom. Mag.*, vol. 22, no. 3, pp. 36–52, Sep. 2015.

[18] R. Gervasi, L. Mastrogiacomo, and F. Franceschini, "A conceptual framework to evaluate human–robot collaboration," *Int. J. Adv. Manuf. Technol.*, vol. 108, no. 3, pp. 841–865, May 2020.

[19] P. Damacharla, A. Y. Javaid, J. J. Gallimore, and V. K. Devabhaktuni, "Common metrics to benchmark human–machine teams (HMT): A review," *IEEE Access*, vol. 6, pp. 38637–38655, 2018.

[20] P. A. Lasota, T. Fong, and J. A. Shah, "A survey of methods for safe human–robot interaction," *Found. Trends Robot.*, vol. 5, no. 3, pp. 261–349, 2017.

[21] J. Nelles, S. T. Kwee-Meier, and A. Mertens, "Evaluation metrics regarding human well-being and system performance in human–robot interaction—A literature review," in *Proc. 20th Congr. Int. Ergonom. Assoc.*, Florence, Italy, 2018, pp. 124–135.

[22] R. R. Murphy and D. Schreckenghost, "Survey of metrics for human–robot interaction," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2013, pp. 197–198.

[23] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common metrics for human–robot interaction," in *Proc. 1st ACM SIGCHI/SIGART Conf. Hum.-Robot Interact.*, Salt Lake City, UT, USA, Mar. 2006, pp. 33–40.

[24] S. Singer and D. Akin, "A survey of quantitative team performance metrics for human–robot collaboration," in *Proc. 41st Int. Conf. Environ. Syst.*, Portland, OR, USA, Jul. 2011, p. 5248.

[25] D. Riedelbauch and J. Hümmer, "A benchmark toolkit for collaborative human–robot interaction," in *Proc. 31st IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2022, pp. 806–813.

[26] N. Höllerich and D. Henrich, "Coloured Petri nets for monitoring human actions in flexible human–robot teams," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 2749–2756.

[27] N. Hollerich and D. Henrich, "Relevant perception modalities for flexible human–robot teams," in *Proc. 29th IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 938–943.

[28] D. Riedelbauch and D. Henrich, "Exploiting a human-aware world model for dynamic task allocation in flexible human–robot teams," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6511–6517.

[29] D. Riedelbauch, S. Schweizer, and D. Henrich, "Skill interaction categories for communication in flexible human–robot teams," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Macau, Nov. 2019, pp. 3810–3816.

[30] D. Riedelbauch and D. Henrich, "Coordinating flexible human–robot teams by local world state observation," in *Proc. 26th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2017, pp. 1000–1005.

[31] M. Lorenzini, M. Lagomarsino, L. Fortini, S. Gholami, and A. Ajoudani, "Ergonomic human–robot collaboration in industry: A review," *Frontiers Robot. AI*, vol. 9, p. 262, Jan. 2023.

[32] A. Castro, F. Silva, and V. Santos, "Trends of human–robot collaboration in industry contexts: Handover, learning, and metrics," *Sensors*, vol. 21, no. 12, p. 4113, Jun. 2021.

[33] S. Kumar, C. Savur, and F. Sahin, "Survey of human–robot collaboration in industrial settings: Awareness, intelligence, and compliance," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 1, pp. 280–297, Jan. 2021.

[34] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Trans. Hum.-Mach. Syst.*, vol. 49, no. 3, pp. 209–218, Jun. 2019.

[35] D. Olsen and M. Goodrich, "Metrics for evaluating human–robot interactions," in *Proc. 4th Int. Workshop Perform. Metrics Intell. Syst. (PERMIS)*, 2003, p. 4.

[36] *Robots and Robotic Devices—Collaborative Robots*, Standard ISO/TS 15066:2016, 2016.

[37] R. G. Freedman, S. J. Levine, B. C. Williams, and S. Zilberstein, "Helpfulness as a key metric of human–robot collaboration," 2020, *arXiv:2010.04914*.

[38] K. Beumelburg, "Fähigkeitsorientierte Montageablaufplanung in der direkten Mensch-Roboter-Kooperation," Ph.D. dissertation, Inst. Ind. Manuf. Manag., Univ. Stuttgart, Stuttgart, Germany, 2005.

[39] J. A. Marvel, R. Bostelman, and J. Falco, "Multi-robot assembly strategies and metrics," *ACM Comput. Surveys*, vol. 51, no. 1, pp. 1–32, Jan. 2018.

[40] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human–robot interaction," *J. Hum.-Robot Interact.*, vol. 3, no. 2, pp. 74–99, 2014.

[41] H. A. Yanco and J. Drury, "Classifying human–robot interaction: An updated taxonomy," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, The Hague, The Netherlands, Jun. 2004, pp. 2841–2846.

[42] J. Steven Moore and A. Garg, "The strain index: A proposed method to analyze jobs for risk of distal upper extremity disorders," *Amer. Ind. Hygiene Assoc. J.*, vol. 56, no. 5, pp. 443–458, May 1995.

[43] S. Hignett and L. McAtamney, "Rapid entire body assessment (REBA)," *Appl. Ergonom.*, vol. 31, no. 2, pp. 201–205, 2000.

[44] L. McAtamney and E. N. Corlett, "RULA: A survey method for the investigation of work-related upper limb disorders," *Appl. Ergonom.*, vol. 24, no. 2, pp. 91–99, 1993.

[45] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Jan. 1988.

[46] J. Brooke, "SUS: A 'quick and dirty' usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996.

[47] C. D. Spielberger, "State-Trait Anxiety Inventory for Adults (STAI-AD)," Amer. Psychol. Assoc., Washington, DC, USA, 1983, doi: 10.1037/t06496-000.

[48] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human–robot collaboration," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.–Robot Interact.*, Mar. 2015, pp. 51–58.

[49] K. Inoue, S. Nonaka, Y. Ujiie, T. Takubo, and T. Arai, "Comparison of human psychology for real and virtual mobile manipulators," in *IEEE Int. Workshop Robot Hum. Interact. Commun. (RO-MAN)*, 2005, pp. 73–78.

[50] G. Charalambous, S. Fletcher, and P. Webb, "The development of a scale to evaluate trust in industrial human–robot collaboration," *Int. J. Social Robot.*, vol. 8, no. 2, pp. 193–209, Apr. 2016.

[51] V. V. Unhelkar, H. C. Siu, and J. A. Shah, "Comparative performance of human and mobile robotic assistants in collaborative fetch-and-deliver tasks," in *Proc. ACM/IEEE Int. Conf. Hum. Robot Interact.*, Mar. 2014, pp. 82–89.

[52] M. Rubagotti, I. Tusseyeva, S. Baltabayeva, D. Summers, and A. Sandygulova, "Perceived safety in physical human–robot interaction—A survey," *Robot. Auto. Syst.*, vol. 151, May 2022, Art. no. 104047.

[53] P. Fratczak, Y. M. Goh, P. Kinnell, A. Soltoggio, and L. Justham, "Understanding human behaviour in industrial human–robot interaction by means of virtual reality," in *Proc. Halfway Future Symp.*, Nov. 2019, pp. 1–7.

[54] Y.-J. Zhang, L. Liu, N. Huang, R. Radwin, and J. Li, "From manual operation to collaborative robot assembly: An integrated model of productivity and ergonomic performance," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 895–902, Apr. 2021.

[55] J. Lindblom and W. Wang, "Towards an evaluation framework of safety, trust, and operator experience in different demonstrators of human–robot collaboration," in *Advances in Manufacturing Technolog* (Advances in Transdisciplinary Engineering). Amsterdam, The Netherlands: IOS Press, 2018, pp. 145–150.

[56] N. Kang, C. Zhao, J. Li, and J. A. Horst, "A hierarchical structure of key performance indicators for operation management and continuous improvement in production systems," *Int. J. Prod. Res.*, vol. 54, no. 21, pp. 6333–6350, Nov. 2016.

[57] H. B. Maynard, G. J. Stegemerten, and J. L. Schwab, *Methods-Time Measurement*. New York, NY, USA: McGraw-Hill, 1948.

[58] A. Meissner, A. Trübswetter, A. S. Conti-Kufner, and J. Schmidtler, "Friend or foe? Understanding assembly workers' acceptance of human–robot collaboration," *ACM Trans. Hum.-Robot Interact.*, vol. 10, no. 1, pp. 1–30, Mar. 2021.

[59] A. Downs, W. Harrison, and C. Schlenoff, "Test methods for robot agility in manufacturing," *Ind. Robot, Int. J.*, vol. 43, no. 5, pp. 563–572, Aug. 2016.

[60] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, Nov. 2018.

[61] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human–robot collaborative tasks," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2015, pp. 189–196.

[62] T. S. Sievers, B. Schmitt, P. Rückert, M. Petersen, and K. Tracht, "Concept of a mixed-reality learning environment for collaborative robotics," *Proc. Manuf.*, vol. 45, pp. 19–24, Jan. 2020.

[63] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[64] M. R. Endsley and D. B. Kaber, "Level of automation effects on performance, situation awareness and workload in a dynamic control task," *Ergonomics*, vol. 42, no. 3, pp. 462–492, Mar. 1999.

[65] C. Juelg, A. Hermann, A. Roennau, and R. Dillmann, "Efficient, collaborative screw assembly in a shared workspace," in *Advances in Intelligent Systems and Computing*, vol. 867. New York, NY, USA: Springer-Verlag, 2019, pp. 837–848.

[66] P. Tsarouchi, A.-S. Matthaiakis, S. Makris, and G. Chryssolouris, "On a human–robot collaboration in an assembly cell," *Int. J. Comput. Integr. Manuf.*, vol. 30, no. 6, pp. 580–589, 2017.

[67] A. Garg, J. S. Moore, and J. M. Kapellusch, "The revised strain index: An improved upper extremity exposure assessment model," *Ergonomics*, vol. 60, no. 7, pp. 912–922, Jul. 2017.

[68] K. Schaub, G. Caragnano, B. Britzke, and R. Bruder, "The European assembly worksheet," *Theor. Issues Ergonom. Sci.*, vol. 14, no. 6, pp. 616–639, Nov. 2013.

[69] T. R. Waters, V. Putz-Anderson, A. Garg, and L. J. Fine, "Revised NIOSH equation for the design and evaluation of manual lifting tasks," *Ergonomics*, vol. 36, no. 7, pp. 749–776, Jul. 1993.

[70] F. Fusaro, E. Lamon, E. D. Momi, and A. Ajoudani, "A human-aware method to plan complex cooperative and autonomous tasks using behavior trees," in *Proc. IEEE-RAS 20th Int. Conf. Humanoid Robots (Humanoids)*, Jul. 2021, pp. 1–10.

[71] E. Lamon, "Ergonomic and worker-centric human–robot collaboration: Strategies, interfaces and controllers," Ph.D. dissertation, Istituto Italiano di Tecnologia, Univ. Pisa, Italy, 2021.

[72] I. E. Makrini, K. Merckaert, J. D. Winter, D. Lefeber, and B. Vanderborght, "Task allocation for improved ergonomics in human–robot collaborative assembly," *Interact. Stud.*, vol. 20, no. 1, pp. 102–133, Jul. 2019.

[73] B. Busch, M. Toussaint, and M. Lopes, "Planning ergonomic sequences of actions in human–robot interaction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1916–1923.

[74] W. Kim, M. Lorenzini, K. Kapicioğlu, and A. Ajoudani, "ErgoTac: A tactile feedback interface for improving human ergonomics in workplaces," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4179–4186, Oct. 2018.

[75] M. Lorenzini, W. Kim, E. D. Momi, and A. Ajoudani, "A new overloading fatigue model for ergonomic risk assessment with application to human–robot collaboration," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 1962–1968.

[76] L. Peternel, C. Fang, N. Tsagarakis, and A. Ajoudani, "A selective muscle fatigue management approach to ergonomic human–robot co-manipulation," *Robot. Comput.-Integr. Manuf.*, vol. 58, pp. 69–79, Aug. 2019.

[77] L. Peternel, C. Fang, N. Tsagarakis, and A. Ajoudani, "Online human muscle force estimation for fatigue management in human–robot co-manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Madrid, Spain, Oct. 2018, pp. 1340–1346.

[78] A. Ranavolo, F. Draicchio, T. Varrecchia, A. Silvetti, and S. Iavicoli, "Wearable monitoring devices for biomechanical risk assessment at work: Current status and future challenges—A systematic review," *Int. J. Environ. Res. Public Health*, vol. 15, no. 9, p. 2001, Sep. 2018.

[79] L. Longo, "Experienced mental workload, perception of usability, their interaction and impact on task performance," *PLoS ONE*, vol. 13, no. 8, Aug. 2018, Art. no. e0199661.

[80] B. Xie and G. Salvendy, "Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments," *Work Stress*, vol. 14, no. 1, pp. 74–99, Jan. 2000.

[81] L. Gualtieri, F. Fraboni, M. D. Marchi, and E. Rauch, "Evaluation of variables of cognitive ergonomics in industrial human–robot collaborative assembly systems," in *Proc. 21st Congr. Int. Ergonom. Assoc. (IEA)*. Cham, Switzerland: Springer, 2021, pp. 266–273.

[82] F. Nachreiner, "Standards for ergonomics principles relating to the design of work systems and to mental workload," *Appl. Ergonom.*, vol. 26, no. 4, pp. 259–263, Aug. 1995.

[83] W. Lambrechts, J. S. Klaver, L. Koudijzer, and J. Semeijn, "Human factors influencing the implementation of cobots in high volume distribution centres," *Logistics*, vol. 5, no. 2, p. 32, May 2021.

[84] S. Zoghbi, E. Croft, D. Kulić, and M. Van der Loos, "Evaluation of affective state estimations using an on-line reporting device during human–robot interactions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 3742–3749.

[85] K. L. Koay, K. Dautenhahn, S. N. Woods, and M. L. Walters, "Empirical results from using a comfort level device in human–robot interaction studies," in *Proc. 1st ACM SIGCHI/SIGART Conf. Hum.-Robot Interact.*, Salt Lake City, UT, USA, Mar. 2006, p. 194.

[86] K. Hald, M. Rehm, and T. B. Moeslund, "Proposing human–robot trust assessment through tracking physical apprehension signals in close-proximity human–robot collaboration," in *Proc. 28th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, New Delhi, India, Oct. 2019, pp. 1–6.

[87] J. Baraglia, M. Cakmak, Y. Nagai, R. P. Rao, and M. Asada, "Efficient human–robot collaboration: When should a robot take initiative?" *Int. J. Robot. Res.*, vol. 36, nos. 5–7, pp. 563–579, 2017.

[88] J. Shah, J. Wiken, B. Williams, and C. Breazeal, "Improved human–robot team performance using chaski, a human-inspired plan execution system," in *Proc. 6th Int. Conf. Hum.-Robot Interact.*, New York, NY, USA, Mar. 2011, pp. 29–36.

[89] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human–robot collaboration," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, New York, NY, USA, Feb. 2018, pp. 307–315.

[90] S. M. M. Rahman and Y. Wang, "Mutual trust-based subtask allocation for human–robot collaboration in flexible lightweight assembly in manufacturing," *Mechatronics*, vol. 54, pp. 94–109, Oct. 2018.

[91] B. Sadrfaridpour, H. Saeidi, J. Burke, K. Madathil, and Y. Wang, "Modeling and control of trust in human–robot collaborative manufacturing," in *Robust Intelligence and Trust in Autonomous Systems*. Cham, Switzerland: Springer, 2016, pp. 115–141.

[92] J. A. Saleh, F. Karray, and M. Morckos, "Modelling of robot attention demand in human–robot interaction using finite fuzzy state automata," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Brisbane, QLD, Australia, Jun. 2012, pp. 1–8.

[93] A. H. Memar and E. T. Esfahani, "Objective assessment of human workload in physical human–robot cooperation using brain monitoring," *ACM Trans. Hum.-Robot Interact.*, vol. 9, no. 2, pp. 1–21, Jun. 2020.

[94] C. Setz, B. Arnrich, J. Schumm, R. L. Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, Mar. 2010.

[95] M. Sorostinean, F. Ferland, and A. Tapus, "Reliable stress measurement using face temperature variation with a thermal camera in human–robot interaction," in *Proc. IEEE-RAS 15th Int. Conf. Humanoid Robots (Humanoids)*, Seoul, South Korea, Nov. 2015, pp. 14–19.

[96] I. Eimontaite, I. Gwilt, D. Cameron, J. M. Aitken, J. Rolph, S. Mokaram, and J. Law, "Language-free graphical signage improves human performance and reduces anxiety when working collaboratively with robots," *Int. J. Adv. Manuf. Technol.*, vol. 100, nos. 1–4, pp. 55–73, Sep. 2018.

[97] S. J. Baltrusch, F. Krause, A. W. de Vries, W. van Dijk, and M. P. de Looze, "What about the human in human robot collaboration?" *Ergonomics*, vol. 65, pp. 1–22, Oct. 2021.

[98] S. Hopko, J. Wang, and R. Mehta, "Human factors considerations and metrics in shared space human–robot collaboration: A systematic review," *Frontiers Robot. AI*, vol. 9, p. 6, Feb. 2022.

[99] D. Wurhofer, V. Fuchsberger, T. Meneweger, C. Moser, and M. Tscheligi, "Insights from user experience research in the factory: What to consider in interaction design," in *Human Work Interaction Design Work Analysis and Interaction Design Methods for Pervasive and Smart Workplaces*. Cham, Switzerland: Springer, 2015, pp. 39–56.

[100] W. S. Helton, G. J. Funke, and B. A. Knott, "Measuring workload in collaborative contexts: Trait versus state perspectives," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 56, no. 2, pp. 322–332, Mar. 2014.

[101] K. Finstad, "The usability metric for user experience," *Interacting Comput.*, vol. 22, no. 5, pp. 323–327, Sep. 2010.

[102] G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: A scaling procedure for measuring mental workload," in *Advances in Psychology*. Amsterdam, The Netherlands: Elsevier, 1988, pp. 185–218.

[103] F. R. H. Zijlstra and L. V. Doorn, *The Construction of a Scale To Measure Subjective Effort*, vol. 43. Delft, Netherlands: Delft Univ. Press, 1985, pp. 124–139.

[104] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Social Robot.*, vol. 1, no. 1, pp. 71–81, Nov. 2008.

[105] E. R. Thompson, "Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS)," *J. Cross-Cultural Psychol.*, vol. 38, no. 2, pp. 227–242, Mar. 2007.

[106] T. Nomura, T. Kanda, and T. Suzuki, "Experimental investigation into influence of negative attitudes toward robots on human–robot interaction," *AI Soc.*, vol. 20, no. 2, pp. 138–150, Mar. 2006.

[107] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, "Measurement of negative attitudes toward robots," *Interact. Stud.*, vol. 7, no. 3, pp. 437–454, Nov. 2006.

[108] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quart.*, vol. 27, pp. 425–478, Sep. 2003.

[109] M. C. Gombolay, R. A. Gutierrez, S. G. Clarke, G. F. Sturla, and J. A. Shah, "Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams," *Auto. Robots*, vol. 39, no. 3, pp. 293–312, Oct. 2015.

[110] M. Gombolay, A. Bair, C. Huang, and J. Shah, "Computational design of mixed-initiative human–robot teaming that considers human factors: Situational awareness, workload, and workflow preferences," *Int. J. Robot. Res.*, vol. 36, nos. 5–7, pp. 597–617, Jun. 2017.

[111] M. Paliga and A. Pollak, "Development and validation of the fluency in human–robot interaction scale. A two-wave study on three perspectives of fluency," *Int. J. Hum.-Comput. Stud.*, vol. 155, Nov. 2021, Art. no. 102698.

[112] C. Bröhl, J. Nelles, C. Brandl, A. Mertens, and C. M. Schlick, "TAM reloaded: A technology acceptance model for human–robot cooperation in production systems," in *Proc. HCI Int. Posters Extended Abstracts, 18th Int. Conf.* Cham, Switzerland: Springer, 2016, pp. 97–103.

[113] C. Bröhl, J. Nelles, C. Brandl, A. Mertens, and V. Nitsch, "Human–robot collaboration acceptance model: Development and comparison for Germany, Japan, China and the USA," *Int. J. Social Robot.*, vol. 11, no. 5, pp. 709–726, Nov. 2019.

[114] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the trust perception scale-HRI," in *Robust Intelligence and Trust in Autonomous Systems*. New York, NY, USA: Springer, 2016, pp. 191–218.

[115] R. Etzi, S. Huang, G. W. Scurati, S. Lyu, F. Ferrise, A. Gallace, A. Gaggioli, A. Chirico, M. Carulli, and M. Bordegoni, "Using virtual reality to test human–robot interaction during a collaborative task," in *Proc. ASME Design Eng. Tech. Conf.*, vol. 1. New York, NY, USA: American Society of Mechanical Engineers (ASME), Nov. 2019, Art. no. V001T02A080.

[116] C. Chao and A. Thomaz, "Timing in multimodal turn-taking interactions: Control and analysis using timed Petri nets," *J. Hum.-Robot Interact.*, vol. 1, no. 1, pp. 4–25, Aug. 2012.

[117] R. Palmarini, I. F. Del Amo, G. Bertolino, G. Dini, J. A. Erkoyuncu, R. Roy, and M. Farnsworth, "Designing an AR interface to improve trust in human–robots collaboration," *Proc. CIRP*, vol. 70, no. 1, pp. 350–355, Jan. 2018.

[118] P. A. Lasota and J. A. Shah, "Analyzing the effects of human-aware motion planning on close-proximity human–robot collaboration," *Hum. Factors, J. Human Factors Ergonom. Soc.*, vol. 57, no. 1, pp. 21–33, Jan. 2015.

[119] M. Lagomarsino, M. Lorenzini, E. De Momi, and A. Ajoudani, "Robot trajectory adaptation to optimise the trade-off between human cognitive ergonomics and workplace productivity in collaborative tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 663–669.

[120] M. Lagomarsino, M. Lorenzini, E. De Momi, and A. Ajoudani, "An online framework for cognitive load assessment in industrial tasks," *Robot. Comput.-Integr. Manuf.*, vol. 78, Dec. 2022, Art. no. 102380.

[121] C. Messeri, G. Masotti, A. M. Zanchettin, and P. Rocco, "Human–robot collaboration: Optimizing stress and productivity based on game theory," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8061–8068, Oct. 2021.

[122] A. Pollak, M. Paliga, M. M. Pulopulos, B. Kozusznik, and M. W. Kozusznik, "Stress in manual and autonomous modes of collaboration with a cobot," *Comput. Hum. Behav.*, vol. 112, Nov. 2020, Art. no. 106469.

[123] R. Kato, M. Fujita, and T. Arai, "Development of advanced cellular manufacturing system with human–robot collaboration," in *Proc. 19th Int. Symp. Robot Hum. Interact. Commun.*, Viareggio, Italy, Sep. 2010, pp. 355–360.

[124] J. Sims, D. Vashishtha, P. Rani, R. Brackin, and N. Sarkar, "Stress detection for implicit human–robot co-operation," in *Proc. 5th Biannual World Autom. Congr.*, 2002, pp. 1–11.

[125] F. Zhao, C. Henrichs, and B. Mutlu, "Task interdependence in human–robot teaming," in *Proc. 29th IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 1143–1149.

[126] D. Kulic and E. A. Croft, "Affective state estimation for human–robot interaction," *IEEE Trans. Robot.*, vol. 23, no. 5, pp. 991–1000, Oct. 2007.

[127] D. Kulic and E. Croft, "Estimating robot induced affective state using hidden Markov models," in *Proc. ROMAN 15th IEEE Int. Symp. Robot Hum. Interact. Commun.*, Hatfield, U.K., Sep. 2006, pp. 257–262.

[128] D. Kulic and E. Croft, "Anxiety detection during human–robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Edmonton, AB, Canada, Aug. 2005, pp. 616–621.

[129] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system-towards implicit human–robot collaboration," *Robotica*, vol. 22, no. 1, pp. 85–95, Jan. 2004.

[130] V. Weistroffer, A. Paljic, P. Fuchs, O. Hugues, J.-P. Chodacki, P. Ligot, and A. Morais, "Assessing the acceptability of human–robot co-presence on assembly lines: A comparison between actual situations and their virtual reality counterparts," in *Proc. 23rd IEEE Int. Symp. Robot Hum. Interact. Commun.*, Edinburgh, Scotland, Aug. 2014, pp. 377–384.

[131] J. Morton, P. Vanneste, C. Larmuseau, B. B. Van Acker, A. Raes, K. Bombeke, F. Cornillie, J. Saldien, and L. De, "Identifying predictive EEG features for cognitive overload detection in assembly workers in industry 4.0," in *Proc. H-Workload 3rd Int. Symp. Hum. Mental Workload, Models Appl. (Works Prog.)*, 2019, p. 1.

[132] J. Coyne and C. Sibley, "Investigating the use of two low cost eye tracking systems for detecting pupillary response to changes in mental workload," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 60, no. 1, pp. 37–41, 2016.

[133] Y. Yamada, Y. Umetani, and Y. Hirasawa, "Proposal of a psychophysiological experiment system applying the reaction of human pupillary dilation to frightening robot motions," in *Proc. IEEE SMC99 Conf. Int. Conf. Syst., Man, Cybern.*, vol. 2, Tokyo, Japan, Mar. 1999, pp. 1052–1057.

[134] Y. Hu, M. Benallegue, G. Venture, and E. Yoshida, "Interact with me: An exploratory study on interaction factors for active physical human–robot interaction," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6764–6771, Oct. 2020.

[135] I. Eimontaite, D. Cameron, J. Rolph, S. Mokaram, J. M. Aitken, I. Gwilt, and J. Law, "Dynamic graphical instructions result in improved attitudes and decreased task completion time in human–robot co-working: An experimental manufacturing study," *Sustainability*, vol. 14, no. 6, p. 3289, Mar. 2022.

[136] G. Tang, P. Webb, and J. Thrower, "The development and evaluation of robot light skin: A novel robot signalling system to improve communication in industrial human–robot collaboration," *Robot. Comput.-Integr. Manuf.*, vol. 56, pp. 85–94, Apr. 2019.

[137] M. Fujita, R. Kato, and A. Tamio, "Assessment of operators' mental strain induced by hand-over motion of industrial robot manipulator," in *Proc. 19th Int. Symp. Robot Hum. Interact. Commun.*, Viareggio, Italy, Sep. 2010, pp. 361–366.

[138] F. Dehais, E. A. Sisbot, R. Alami, and M. Causse, "Physiological and subjective evaluation of a human–robot object hand-over task," *Appl. Ergonom.*, vol. 42, no. 6, pp. 785–791, 2011.

[139] N. Hanajima, Y. Ohta, Y. Sakurai, H. Hikita, and M. Yamashita, "Further experiments to investigate the influence of robot motions on human impressions," in *Proc. 15th IEEE Int. Symp. Robot Human Interact. Commun.*, Hatfield, U.K., Sep. 2006, pp. 733–740.

[140] N. Hanajima, M. Fujimoto, H. Hikita, and M. Yamashita, "Influence of auditory and visual modalities on skin potential response to robot motions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sendai, Jun. Japan, 2004, pp. 1226–1231.

[141] M. Grassmann, E. Vlemincx, A. Von Leupoldt, J. M. Mittelstädt, and O. Van Den Bergh, "Respiratory changes in response to cognitive load: A systematic review," *Neural Plasticity*, vol. 2016, pp. 1–16, May 2016.

[142] I. B. Ajenaghughrure, S. D. C. Sousa, and D. Lamas, "Measuring trust with psychophysiological signals: A systematic mapping study of approaches used," *Multimodal Technol. Interact.*, vol. 4, no. 3, p. 63, Sep. 2020.

[143] E. Debie, R. Fernandez Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal fusion for objective assessment of cognitive workload: A review," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1542–1555, Mar. 2021.

[144] P. Vanneste, A. Raes, J. Morton, K. Bombeke, B. B. V. Acker, C. Larmuseau, F. Depaepe, and W. V. Den Noortgate, "Towards measuring cognitive load through multimodal physiological data," *Cognition, Technol. Work*, vol. 23, pp. 567–585, Jul. 2020.

[145] J. Heard, R. Heald, C. E. Harriott, and J. A. Adams, "A diagnostic human workload assessment algorithm for collaborative and supervisory human–robot teams," *ACM Trans. Hum.-Robot Interact.*, vol. 8, no. 2, pp. 1–30, Jun. 2019.

[146] D. Novak, M. Mihelj, and M. Munih, "Psychophysiological responses to different levels of cognitive and physical workload in haptic interaction," *Robotica*, vol. 29, no. 3, pp. 367–374, May 2010.

[147] L. Tiberio, A. Cesta, and M. O. Belardinelli, "Psychophysiological methods to evaluate user's response in human robot interaction: A review and feasibility study," *Robotics*, vol. 2, no. 2, pp. 92–121, Jun. 2013.

[148] C. L. Bethel, K. Salomon, R. R. Murphy, and J. L. Burke, "Survey of psychophysiology measurements applied to human–robot interaction," in *Proc. 16th IEEE Int. Symp. Robot Hum. Interact. Commun.*, 2007, pp. 732–737.

[149] T. Zhang, J. Yang, N. Liang, B. J. Pitts, K. O. Prakah-Asante, R. Curry, B. S. Duerstock, J. P. Wachs, and D. Yu, "Physiological measurements of situation awareness: A systematic review," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 2020, Nov. 2020, Art. no. 001872082096907.

[150] A. G. Sartang, M. Ashnagar, E. Habibi, and S. Sadeghi, "Evaluation of rating scale mental effort (RSME) effectiveness for mental workload assessment in nurses," *J. Occupational Health Epidemiol.*, vol. 5, no. 4, pp. 211–217, Oct. 2016.

[151] M. A. Hogg and D. Abrams, "Social cognition and attitudes," in *Psychology*, 3rd ed., G. N. Martin, N. R. Carlson, and W. Buskist, Eds. London, U.K.: Pearson Education Limited, 2007, pp. 684–721.

[152] *Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts*, ISO Standard 9241-11:2018, 2018.

[153] J. R. Lewis and J. Sauro, "The factor structure of the system usability scale," in *Human Centered Design*. Berlin, Germany: Springer, 2009, pp. 94–103.

[154] J. Sauro, *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. Denver, CO, USA: Measuring Usability LLC, 2011.

[155] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models," *Manage. Sci.*, vol. 35, pp. 982–1003, Aug. 1989.

[156] V. Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions," *Decis. Sci.*, vol. 39, no. 2, pp. 273–315, 2008.

[157] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology," *MIS Quart.*, vol. 36, no. 1, pp. 157–178, Feb. 2012.

[158] R. Brauer, "Acceptance of cooperative robots in the industrial context," Ph.D. dissertation, Univ. Technol. Chemnitz, Chemnitz, Germany, 2017.

[159] M. Ghazizadeh, J. D. Lee, and L. N. Boyle, "Extending the technology acceptance model to assess automation," *Cognition, Technol. Work*, vol. 14, no. 1, pp. 39–49, Oct. 2011.

[160] G. Hoffman, "Evaluating fluency in human–robot collaboration," *Proc. HRI Workshop Hum. Robot Collaboration*, 2013, pp. 209–218.

[161] S. Nikolaidis, D. Hsu, and S. Srinivasa, "Human–robot mutual adaptation in collaborative tasks: Models and experiments," *Int. J. Robot. Res.*, vol. 36, nos. 5–7, pp. 618–634, Feb. 2017.

[162] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 46, no. 1, pp. 50–80, Mar. 2004.

[163] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognit. Ergonom.*, vol. 4, no. 1, pp. 53–71, Mar. 2000.

[164] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, Mar. 1996.

[165] T. T. Kessler, C. Larios, T. Walker, V. Yerdon, and P. A. Hancock, "A comparison of trust measures in human–robot interaction scenarios," in *Advances in Human Factors in Robots and Unmanned Systems*. Cham, Switzerland: Springer, Jul. 2016, pp. 353–364.

[166] V. Groom and C. Nass, "Can robots be teammates?: Benchmarks in human–robot teams," *Interact. Stud.*, vol. 8, no. 3, pp. 483–500, Nov. 2007.

[167] Z. R. Khavas, S. R. Ahmadzadeh, and P. Robinette, "Modeling trust in human–robot interaction: A survey," in *Proc. Int. Conf. Social Robot.* Cham, Switzerland: Springer, 2020, pp. 529–541.

[168] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human–robot interaction," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 53, no. 5, pp. 517–527, 2011.

[169] M. Endsley, "Situation awareness global assessment technique (SAGAT)," in *Proc. IEEE Nat. Aerosp. Electron. Conf.*, Dec. 1988, pp. 789–795.

[170] F. Caputo, A. Greco, E. D'Amato, I. Notaro, and S. Spada, "IMU-based motion capture wearable system for ergonomic assessment in industrial environment," in *Advances in Human Factors in Wearable Technologies and Game Design*. Cham, Switzerland: Springer, 2018, pp. 215–225.

[171] A. G. Marin, M. S. Shourijeh, P. E. Galibarov, M. Damsgaard, L. Fritzsch, and F. Stulp, "Optimizing contextual ergonomics models in human–robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–9.

[172] N. Vignais, M. Miezal, G. Bleser, K. Mura, D. Gorecky, and F. Marin, "Innovative system for real-time ergonomic feedback in industrial manufacturing," *Appl. Ergonom.*, vol. 44, no. 4, pp. 566–574, Jul. 2013.

[173] W. Kim, J. Lee, N. Tsagarakis, and A. Ajoudani, "A real-time and reduced-complexity approach to the detection and monitoring of static joint overloading in humans," in *Proc. Int. Conf. Rehabil. Robot. (ICORR)*, Jul. 2017, pp. 1–9.

[174] K. P. Hawkins, S. Bansal, N. N. Vo, and A. F. Bobick, "Anticipating human actions for collaboration in the presence of task and sensor uncertainty," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Hong Kong, May 2014, pp. 2215–2222.

[175] L. Wang, R. Gao, J. Váncza, J. Krüger, X. V. Wang, S. Makris, and G. Chryssolouris, "Symbiotic human–robot collaborative assembly," *CIRP Ann.*, vol. 68, no. 2, pp. 701–726, 2019.

[176] A. M. Zanchettin, A. Casalino, L. Piroddi, and P. Rocco, "Prediction of human activity patterns for human–robot collaborative assembly tasks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 3934–3942, Jul. 2019.

[177] A. Ghadirzadeh, X. Chen, W. Yin, Z. Yi, M. Bjorkman, and D. Kragic, "Human-centered collaborative robots with deep reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 566–571, Apr. 2021.

[178] O. C. Schrempf, D. Albrecht, and U. D. Hanebeck, "Tractable probabilistic models for intention recognition based on expert knowledge," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, San Diego, CA, USA, Oct. 2007, pp. 1429–1434.

[179] P. Kratzer, S. Bihlmaier, N. B. Midlagajni, R. Prakash, M. Toussaint, and J. Mainprice, "MoGaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 367–373, Apr. 2021.

[180] P. A. Lasota, G. F. Rossano, and J. A. Shah, "Toward safe close-proximity human–robot interaction with standard industrial robots," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, New Taipei, Taiwan, Aug. 2014, pp. 339–344.

[181] G. Nagymáté and R. M. Kiss, "Application of OptiTrack motion capture systems in human movement analysis," *Recent Innov. Mechatronics*, vol. 5, no. 1, pp. 1–9, Jan. 1970.

[182] H. Hu, Z. Cao, X. Yang, H. Xiong, and Y. Lou, "Performance evaluation of optical motion capture sensors for assembly motion capturing," *IEEE Access*, vol. 9, pp. 61444–61454, 2021.

[183] A. Chatzitofis, D. Zarpalas, P. Daras, and S. Kollias, "DeMoCap: Low-cost marker-based motion capture," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3338–3366, Dec. 2021.

[184] J. A. Marvel, "Performance metrics of speed and separation monitoring in shared workspaces," *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 2, pp. 405–414, Apr. 2013.

[185] J. Wallström and J. Lindblom, "Design and development of the USUS goals evaluation framework," in *Human-Robot Interaction: Evaluation Methods and Their Standardization*, vol. 12. Cham, Switzerland: Springer, 2020, pp. 177–201.

[186] M. Rueben, S. A. Elprama, D. Chrysostomou, and A. Jacobs, "Introduction to (Re) using questionnaires in human–robot interaction research," in *Human-Robot Interaction* (Springer Series on Bio- and Neurosystems), vol. 12. Cham, Switzerland: Springer, 2020, pp. 125–144.

[187] *Standards for Educational and Psychological Testing*. Washington, DC, USA: American Educational Research Association, 2014.

[188] E. Cho, "Making reliability reliable," *Organizational Res. Methods*, vol. 19, no. 4, pp. 651–682, Jul. 2016.

[189] R. J. Fisher, "Social desirability bias and the validity of indirect questioning," *J. Consum. Res.*, vol. 20, no. 2, p. 303, Sep. 1993.

[190] A. M. Rosenthal-von der Putten, F. P. Schulte, S. C. Eimler, L. Hoffmann, S. Sobieraj, S. Maderwald, N. C. Kramer, and M. Brand, "Neural correlates of empathy towards robots," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Tokyo, Japan, Mar. 2013, pp. 215–216.

[191] R. Rosenthal and R. L. Rosnow, *Essentials of Behavioral Research: Methods and Data Analysis*, 3rd ed., M. Beth, Ed. New York, NY, USA: McGraw-Hill, 2008.

[192] M. Umair, N. Chalabianloo, C. Sas, and C. Ersoy, "HRV and stress: A mixed-methods approach for comparison of wearable heart rate sensors for biofeedback," *IEEE Access*, vol. 9, pp. 14005–14024, 2021.

[193] P. Mijović, V. Ković, M. D. Vos, I. Mačužić, P. Todorović, B. Jeremić, and I. Gligorijević, "Towards continuous and real-time attention monitoring at work: Reaction time versus brain response," *Ergonomics*, vol. 60, no. 2, pp. 241–254, Feb. 2017.

[194] C. L. Bethel and R. R. Murphy, "Review of human studies methods in HRI and recommendations," *Int. J. Social Robot.*, vol. 2, no. 4, pp. 347–359, 2010.

[195] V. A. Thurmond, "The point of triangulation," *J. Nursing Scholarship*, vol. 33, no. 3, pp. 253–258, Sep. 2001.

[196] M. Koppenborg, P. Nickel, B. Naber, A. Lungfiel, and M. Huelke, "Effects of movement speed and predictability in human–robot collaboration," *Hum. Factors Ergonom. Manuf. Service Industries*, vol. 27, no. 4, pp. 197–209, Jul. 2017.

[197] V. Weistroffer, A. Paljic, L. Callebert, and P. Fuchs, "A methodology to assess the acceptability of human–robot collaboration using virtual reality," in *Proc. 19th ACM Symp. Virtual Reality Softw. Technol.*, Singapore: ACM Press, Oct. 2013, pp. 39–48.

[198] P. Fratczak, Y. M. Goh, P. Kinnell, L. Justham, and A. Soltoggio, "Virtual reality study of human adaptability in industrial human–robot collaboration," in *Proc. IEEE Int. Conf. Hum.-Mach. Syst. (ICHMS)*, Rome, Italy, Sep. 2020, pp. 1–6.

[199] J. Moultrie, "Understanding and classifying the role of design demonstrators in scientific exploration," *Technovation*, vols. 43–44, pp. 1–16, Sep. 2015.

[200] P. Gustavsson, A. Syberfeldt, R. Brewster, and L. Wang, "Human–robot collaboration demonstrator combining speech recognition and haptic control," *Proc. CIRP*, vol. 63, pp. 396–401, Oct. 2017.

[201] P. Rueckert, S. Muenkewarf, and K. Tracht, "Human-in-the-loop simulation for virtual commissioning of human–robot-collaboration," *Proc. CIRP*, vol. 88, pp. 229–233, Nov. 2020.

[202] A. Antakli, T. Spieldenner, M. Köster, J. Groß, E. Herrmann, D. Rubinstein, D. Spieldenner, and I. Zinnikus, "Optimized coordination and simulation for industrial human robot collaborations," in *Proc. Web Inf. Syst. Technol.*, Vienna, Austria, 2019, pp. 44–68.

[203] F. Ore, B. Vemula, L. Hanson, M. Wiktorsson, and B. Fagerström, "Simulation methodology for performance and safety evaluation of human–industrial robot collaboration workstation design," *Int. J. Intell. Robot. Appl.*, vol. 3, no. 3, pp. 269–282, Sep. 2019.

[204] V. Gopinath, K. Johansen, and M. Derelöv, "Demonstrators to support research in industrial safety – a methodology," *Proc. Manuf.*, vol. 17, pp. 246–253, Jan. 2018.

[205] M. Holm, R. Senington, W. Wang, and J. Lindblom, "Real-world industrial demonstrators on humanbrobot collaborative assembly," in *Advanced Human-Robot Collaboration in Manufacturing*. Cham, Switzerland: Springer, 2021, pp. 413–438.

[206] H. Karami, A. Carfi, and F. Mastrogiovanni, "Branched AND/OR graphs: Toward flexible and adaptable human–robot collaboration," in *Proc. 30th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2021, pp. 527–533.

[207] K. Darvish, B. Bruno, E. Simetti, F. Mastrogiovanni, and G. Casalino, "Interleaved online task planning, simulation, task allocation and motion control for flexible human–robot cooperation," in *Proc. 27th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Nanjing, China, Aug. 2018, pp. 58–65.

[208] D. Riedelbauch, T. Werner, and D. Henrich, "Supporting a human-aware world model through sensor fusion," in *Proc. Int. Conf. Robot. Alpe-Adria-Danube Region (RAAD)*, 2017, pp. 665–672.

[209] C. Bartneck, T. Belpaeme, F. Eyssel, T. Kanda, M. Keijsers, and S. Šabanović, "Research methods," in *Human-Robot Interaction—An Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2020, pp. 126–160.

[210] G. Hoffman and X. Zhao, "A primer for conducting experiments in human–robot interaction," *ACM Trans. Hum.-Robot Interact.*, vol. 10, no. 1, pp. 1–31, Mar. 2021.

[211] C. L. Bethel, Z. Henkel, and K. Baugus, "Conducting studies in human–robot interaction," in *Human-Robot Interaction—Evaluation Methods and Their Standardization*. Cham, Switzerland: Springer, 2020, pp. 91–124.

[212] L. Veling and C. McGinn, "Qualitative research in HRI: A review and taxonomy," *Int. J. Social Robot.*, vol. 13, pp. 1–21, Feb. 2021.

[213] K. S. Welfare, M. R. Hallowell, J. A. Shah, and L. D. Riek, "Consider the human work experience when integrating robotics in the workplace," in *Proc. 14th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2019, pp. 75–84.

[214] C. Gaede, F. Ranz, V. Hummerl, and W. Echelmeyer, "A study on challenges in the implementation of human–robot collaboration," *J. Eng., Manage. Oper.*, vol. 1, pp. 29–39, Sep. 2019.

[215] C. T. Landi, V. Villani, F. Ferraguti, L. Sabattini, C. Secchi, and C. Fantuzzi, "Relieving operators' workload: Towards affective robotics in industrial scenarios," *Mechatronics*, vol. 54, pp. 144–154, Oct. 2018.

[216] G. Hoffman and C. Breazeal, "Collaboration in human–robot teams," in *Proc. AIAA 1st Intell. Syst. Tech. Conf.*, Sep. 2004, p. 6434.

[217] C. Messeri, A. Bicchi, A. M. Zanchettin, and P. Rocco, "A dynamic task allocation strategy to mitigate the human physical fatigue in collaborative robotics," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2178–2185, Apr. 2022.

[218] K. Hald, M. Rehm, and T. B. Moeslund, "Human–robot trust assessment using top-down visual tracking after robot task execution mistakes," in *Proc. 30th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Vancouver, BC, Canada, Aug. 2021, pp. 892–898.

[219] S. L. Müller-Abdelrazeq, K. Schönefeld, M. Haberstroh, and F. Hees, "Interacting with collaborative robots—A study on attitudes and acceptance in industrial contexts," in *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction*. Cham, Switzerland: Springer, 2019, pp. 101–117.

[220] K. Fischer, L. C. Jensen, F. Kirstein, S. Stabinger, O. Erkent, D. Shukla, and J. Piater, "The effects of social gaze in human–robot collaborative assembly," in *Proc. Int. Conf. Social Robot. (ICSR)*, Paris, France, 2015, pp. 204–213.

[221] S. Nikolaidis and J. Shah, "Human–robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Tokyo, Japan, Mar. 2013, pp. 33–40.

[222] A. Chacón, P. Ponsa, and C. Angulo, "Cognitive interaction analysis in human–robot collaboration using an assembly task," *Electronics*, vol. 10, no. 11, p. 1317, May 2021.

[223] S. Bansal, R. Newbury, W. Chan, A. Cosgun, A. Allen, D. Kulic, T. Drummond, and C. Isbell, "Supportive actions for manipulation in human–robot coworker teams," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 11261–11267.

[224] B. Kühnlenz, M. Erhart, M. Kainert, Z.-Q. Wang, J. Wilm, and K. Kühnlenz, "Impact of trajectory profiles on user stress in close human–robot interaction," *At-Automatisierungstechnik*, vol. 66, no. 6, pp. 483–491, Jun. 2018.

[225] B. Sadrfaridpour and Y. Wang, "Collaborative assembly in hybrid manufacturing cells: An integrated framework for human–robot interaction," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 3, pp. 1178–1192, Jul. 2018.

[226] H. Zhu, V. Gabler, and D. Wollherr, "Legible action selection in human–robot collaboration," in *Proc. 26th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2017, pp. 354–359.

[227] O. Hugues, V. Weistroffer, A. Paljic, P. Fuchs, A. A. Karim, T. Gaudin, and A. Buendia, "Determining the important subjective criteria in the perception of human-like robot movements using virtual reality," *Int. J. Humanoid Robot.*, vol. 13, no. 2, Jun. 2016, Art. no. 1550033.

[228] D. Bortot, H. Ding, A. Antonopolous, and K. Bengler, "Human motion behavior while interacting with an industrial robot," *Work*, vol. 41, pp. 1699–1707, Jan. 2012.

[229] L. Onnasch and C. L. Hildebrandt, "Impact of anthropomorphic robot design on trust and attention in industrial human–robot interaction," *ACM Trans. Hum.-Robot Interact.*, vol. 11, no. 1, pp. 1–24, Oct. 2021.

[230] F. Legler, D. Langer, F. Dittrich, and A. C. Bullinger, "I don't care what the robot does! Trust in automation when working with a heavy—Load robot," in *Proc. Human Factors Ergonom. Soc. Eur. Annu. Conf.*, 2020, pp. 239–253.

[231] J. E. Michaelis, A. Siebert-Evenstone, D. W. Shaffer, and B. Mutlu, "Collaborative or simply uncaged? Understanding human–robot interactions in automation," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Honolulu, HI, USA: ACM, Apr. 2020, pp. 1–12.

[232] M. Bender, M. Braun, P. Rally, and O. Scholtz, "Lightweight robots in manual assembly—Best to start simply!" Fraunhofer Institute for Industrial Eng. IAO, Stuttgart, Germany, Tech. Rep., 2016. [Online]. Available: https://www.researchgate.net/publication/327744724_Lightweight_robots_in_manual_assembly_-_best_to_start_simply_Examining_companies'_initial_experiences_with_lightweight_robots/citation/download

[233] A. Hietanen, R. Pieters, M. Lanz, J. Latokartano, and J.-K. Kämäräinen, "AR-based interaction for human–robot collaborative manufacturing," *Robot. Comput.-Integr. Manuf.*, vol. 63, Jun. 2020, Art. no. 101891.

[234] F. Babel, J. Kraus, P. Hock, H. Asenbauer, and M. Baumann, "Investigating the validity of online robot evaluations: Comparison of findings from an one-sample online and laboratory study," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, New York, NY, USA, Mar. 2021, pp. 116–120.

[235] T. Inamura and Y. Mizuchi, "SIGVerse: A cloud-based VR platform for research on multimodal human–robot interaction," *Frontiers Robot. AI*, vol. 8, May 2021, Art. no. 549360.

[236] S. Chernova, N. DePalma, E. Morant, and C. Breazeal, "Crowdsourcing human–robot interaction: Application from virtual to physical worlds," in *Proc. RO-MAN*, Jul. 2011, pp. 21–26.

[237] C. G. Lee and S. C. Park, "Survey on the virtual commissioning of manufacturing systems," *J. Comput. Des. Eng.*, vol. 1, no. 3, pp. 213–222, 2014.

[238] T. Lechler, E. Fischer, M. Metzner, A. Mayr, and J. Franke, "Virtual commissioning—Scientific review and exploratory use cases in advanced production systems," *Proc. CIRP*, vol. 81, pp. 1125–1130, Jan. 2019.

[239] E. Eros, M. Dahl, A. Hanna, A. Albo, P. Falkman, and K. Bengtsson, "Integrated virtual commissioning of a ROS2-based collaborative and intelligent automation system," in *Proc. 24th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Zaragoza, Spain, Sep. 2019, pp. 407–413.

[240] A. Arntz, S. C. Eimler, and H. U. Hoppe, "A virtual sandbox approach to studying the effect of augmented communication on human–robot collaboration," *Frontiers Robot. AI*, vol. 8, p. 318, Oct. 2021.

[241] A. Dimitrokalli, G.-C. Vosniakos, D. Nathanael, and E. Matsas, "On the assessment of human–robot collaboration in mechanical product assembly by use of virtual reality," *Proc. Manuf.*, vol. 51, pp. 627–634, Jan. 2020.

[242] J. O. Oyekan, W. Hutabarat, A. Tiwari, R. Grech, M. H. Aung, M. P. Mariani, L. López-Dávalos, T. Ricaud, S. Singh, and C. Dupuis, "The effectiveness of virtual environments in developing collaborative strategies between industrial robots and humans," *Robot. Comput.-Integr. Manuf.*, vol. 55, pp. 41–54, Feb. 2019.

[243] G.-C. Vosniakos, L. Ouillon, and E. Matsas, "Exploration of two safety strategies in human–robot collaborative manufacturing using virtual reality," *Proc. Manuf.*, vol. 38, pp. 524–531, Jun. 2019.

[244] U. Dombrowski, T. Stefanak, and J. Perret, "Interactive simulation of human–robot collaboration using a force feedback device," *Proc. Manuf.*, vol. 11, pp. 124–131, Sep. 2017.

[245] J. Höcherl, A. Adam, T. Schlegl, and B. Wrede, "Human–robot assembly: Methodical design and assessment of an immersive virtual environment for real-world significance," in *Proc. 25th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Sep. 2020, pp. 549–556.

[246] R. Matias and P. Menezes, "A VR application for the analysis of human responses to collaborative robots," in *Proc. 17th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2022, pp. 68–78.

[247] A. Arntz, S. C. Eimler, and H. U. Hoppe, "Augmenting the human–robot communication channel in shared task environments," in *Proc. Int. Conf. Collaboration Technol. Social Comput.*, Sep. 2020, pp. 20–34.

[248] A. Arntz, S. C. Eimler, and H. U. Hoppe, "'The robot-arm talks back to me'—Human perception of augmented human–robot collaboration in virtual reality," in *Proc. IEEE Int. Conf. Artif. Intell. Virtual Reality (AIVR)*, Utrecht, The Netherlands, Dec. 2020, pp. 307–312.

[249] W. Zhu, X. Fan, and Y. Zhang, "Applications and research trends of digital human models in the manufacturing industry," *Virtual Reality Intell. Hardw.*, vol. 1, no. 6, pp. 558–579, Dec. 2019.

[250] L. Fritzsche, R. Jendrusch, W. Leidholdt, S. Bauer, T. Jäckel, and A. Pirger, "Introducing ema (editor for manual work activities)—A new tool for enhancing accuracy and efficiency of human simulations in digital production planning," in *Proc. Int. Conf. Digit. Hum. Model.*, 2011, pp. 272–281.

[251] P. R. Castro, D. Högberg, H. Ramsen, J. Bjursten, and L. Hanson, "Virtual simulation of human–robot collaboration workstations," in *Proc. Congr. Int. Ergonom. Assoc.*, Florence, Italy, Aug. 2018, pp. 250–261.

[252] T. Bänziger, A. Kunz, and K. Wegener, "Optimizing human–robot task allocation using a simulation tool based on standardized work descriptions," *J. Intell. Manuf.*, vol. 31, no. 7, pp. 1635–1648, Oct. 2020.

[253] G. Boschetti, M. Bottin, M. Faccio, and R. Minto, "Multi-robot multi-operator collaborative assembly systems: A performance evaluation model," *J. Intell. Manuf.*, vol. 32, no. 5, pp. 1455–1470, Jan. 2021.

[254] M. Faccio, R. Minto, G. Rosati, and M. Bottin, "The influence of the product characteristics on human–robot collaboration: A model for the performance of collaborative robotic assembly," *Int. J. Adv. Manuf. Technol.*, vol. 106, nos. 5–6, pp. 2317–2331, Jan. 2020.

[255] U. Dombrowski, T. Stefanak, and A. Reimer, "Simulation of human–robot collaboration by means of power and force limiting," *Proc. Manuf.*, vol. 17, pp. 134–141, Jan. 2018.

[256] Z. S. Givi, M. Y. Jaber, and W. P. Neumann, "Modelling worker reliability with learning and fatigue," *Appl. Math. Model.*, vol. 39, no. 17, pp. 5186–5199, Sep. 2015.

[257] K. M. Rabby, M. Khan, A. Karimoddini, and S. X. Jiang, "An effective model for human cognitive performance within a human–robot collaboration framework," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3872–3877.

[258] M. K. Monir Rabby, M. Altaf Khan, A. Karimoddini, and S. X. Jiang, "Modeling of trust within a human–robot collaboration framework," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Toronto, ON, Canada, Oct. 2020, pp. 4267–4272.

[259] E. Billing, E. Bampouni, and M. Lamb, "Automatic selection of viewpoint for digital human modelling," in *Proc. Int. Digit. Hum. Model. Symp. (DHM)*, Skövde, 2020, pp. 61–70.

[260] M. P. Mayer, B. Odenthal, M. Faber, C. Winkelholz, and C. M. Schlick, "Cognitive engineering of automated assembly processes," *Hum. Factors Ergonom. Manuf. Service Industries*, vol. 24, no. 3, pp. 348–368, May 2014.

[261] O. C. Görür, B. Rosman, F. Sivrikaya, and S. Albayrak, "FABRIC: A framework for the design and evaluation of collaborative robots with extended human adaptation," *ACM Trans. Hum.-Robot Interact.*, early access, Mar. 2023, doi: 10.1145/3585276.

[262] D. Riedelbauch, D. Luthardt-Bergmann, and D. Henrich, "A cognitive human model for virtual commissioning of dynamic human–robot teams," in *Proc. 5th IEEE Int. Conf. Robotic Comput. (IRC)*, Nov. 2021, pp. 27–34.

[263] Y. Cohen, S. Shoval, M. Faccio, and R. Minto, "Deploying cobots in collaborative systems: Major considerations and productivity analysis," *Int. J. Prod. Res.*, vol. 60, no. 6, pp. 1815–1831, Mar. 2022.

[264] R. Calvo and P. Gil, "Evaluation of collaborative robot sustainable integration in manufacturing assembly by using process time savings," *Materials*, vol. 15, no. 2, p. 611, Jan. 2022.

[265] M. Faccio, M. Bottin, and G. Rosati, "Collaborative and traditional robotic assembly: A comparison model," *Int. J. Adv. Manuf. Technol.*, vol. 102, nos. 5–8, pp. 1355–1372, Jan. 2019.

[266] C. Ferreira, G. Figueira, and P. Amorim, "Scheduling human–robot teams in collaborative working cells," *Int. J. Prod. Econ.*, vol. 235, May 2021, Art. no. 108094.

[267] N. Chen, N. Huang, R. Radwin, and J. Li, "Analysis of assembly-time performance (ATP) in manufacturing operations with collaborative robots: A systems approach," *Int. J. Prod. Res.*, vol. 60, no. 1, pp. 277–296, Jan. 2022.

[268] Y. J. Zhang, N. Huang, R. G. Radwin, Z. Wang, and J. Li, "Flow time in a human–robot collaborative assembly process: Performance evaluation, system properties, and a case study," *IISE Trans.*, vol. 54, no. 3, pp. 238–250, 2022.

[269] W. Dai and D. Berleant, "Benchmarking contemporary deep learning hardware and frameworks: A survey of qualitative metrics," in *Proc. IEEE 1st Int. Conf. Cognit. Mach. Intell. (CogMI)*, Dec. 2019, pp. 148–155.

[270] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme, "From characterising three years of HRI to methodology and reporting recommendations," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2016, pp. 391–398.

[271] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of Likert scales in human–robot interaction studies," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, New York, NY, USA, Mar. 2020, pp. 43–52.

[272] T. Belpaeme, "Advice to new human–robot interaction researchers," in *Human-Robot Interaction—Evaluation Methods and Their Standardization*. Cham, Switzerland: Springer, 2020, pp. 355–369.

[273] L. Wijnen, P. Bremner, S. Lemaignan, and M. Giuliani, "Performing human–robot interaction user studies in virtual reality," in *Proc. 29th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Naples: Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 794–799.

[274] R. Skarbez, F. P. Brooks, and M. C. Whitton, "A survey of presence and related concepts," *ACM Comput. Surveys*, vol. 50, no. 6, pp. 1–39, Nov. 2018.

[275] M. Slater and M. V. Sanchez-Vives, "Enhancing our lives with immersive virtual reality," *Frontiers Robot. AI*, vol. 3, p. 74, Dec. 2016.

[276] E. Matsas, G.-C. Vosniakos, and D. Batras, "Effectiveness and acceptability of a virtual environment for assessing human–robot collaboration in manufacturing," *Int. J. Adv. Manuf. Technol.*, vol. 92, nos. 9–12, pp. 3903–3917, May 2017.

[277] V. Schwind, P. Knierim, N. Haas, and N. Henze, "Using presence questionnaires in virtual reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, p. 360.

[278] D. J. Harris, J. M. Bird, P. A. Smart, M. R. Wilson, and S. J. Vine, "A framework for the testing and validation of simulated environments in experimentation and training," *Frontiers Psychol.*, vol. 11, p. 605, Mar. 2020.

[279] A. Antakli, E. Hermann, I. Zinnikus, H. Du, and K. Fischer, "Intelligent distributed human motion simulation in human–robot collaboration environments," in *Proc. 18th Int. Conf. Intell. Virtual Agents*, Sydney, NSW, Australia, Nov. 2018, pp. 319–324.

[280] F. Bonsignorio and A. P. Del Pobil, "Toward replicable and measurable robotics research [from the guest editors]," *IEEE Robot. Autom. Mag.*, vol. 22, no. 3, pp. 32–35, Sep. 2015.

[281] J. Tani, A. F. Daniele, G. Bernasconi, A. Camus, A. Petrov, A. Courchesne, B. Mehta, R. Suri, T. Zaluska, M. R. Walter, E. Frazzoli, L. Paull, and A. Censi, "Integrated benchmarking and design for reproducible and accessible evaluation of robotic agents," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 6229–6236.

[282] F. Bonsignorio, "A new kind of article for reproducible research in intelligent robotics [from the field]," *IEEE Robot. Autom. Mag.*, vol. 24, no. 3, pp. 178–182, Sep. 2017.

[283] V. Rajendran, P. Carreno-Medrano, W. Fisher, A. Werner, and D. Kulić, "A framework for human–robot interaction user studies," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 6215–6222.

[284] P. Melo, R. Arrais, and G. Veiga, "Development and deployment of complex robotic applications using containerized infrastructures," in *Proc. IEEE 19th Int. Conf. Ind. Informat. (INDIN)*, Palma de Mallorca: Institute of Electrical and Electronics Engineers Inc., Jul. 2021, pp. 1–10.

[285] A. Portner, M. Hoffmann, S. Zug, and M. Knig, "SwarmRob: A docker-based toolkit for reproducibility and sharing of experimental artifacts in robotics research," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 325–332.

[286] D. Alexandrovsky, S. Putze, M. Bonfert, S. Höffner, P. Michelmann, D. Wenig, R. Malaka, and J. D. Smeddinck, "Examining design choices of questionnaires in VR user studies," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–21.

[287] M. Mara, H. Pichler, S. Gross, K. Meyer, R. Haring, B. Reiterer, M. Heiml, B. Krenn, and T. Layer-Wagner, "CoBot studio VR: A virtual reality game environment for transdisciplinary research on interpretability and trust in human–robot collaboration," in *Proc. Int. Workshop Virtual, Augmented, Mixed Reality HRI (VAM-HRI)*, Boulder, 2021, pp. 1–11.

[288] J. Weisz, Y. Huang, F. Lier, S. Sethumadhavan, and P. Allen, "RoboBench: Towards sustainable robotics system benchmarking," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2016, pp. 3383–3389.

[289] B. Lesage and R. Alexander, "SASSI: Safety analysis using simulation-based situation coverage for Cobot systems," in *Proc. Int. Conf. Comput. Saf., Rel., Secur.* Cham, Switzerland: Springer, Sep. 2021, pp. 195–209.

[290] E. Billing, L. Hanson, M. Lamb, and D. Högberg, "Digital human modelling in action," in *Proc. 15th SweCog Conf.*, 2019, pp. 25–28.

[291] A. Steinfeld, O. C. Jenkins, and B. Scassellati, "The oz of wizard," in *Proc. ACM/IEEE Int. Conf. Hum. Robot Interact. (HRI)*, Dec. 2009, pp. 101–108.

[292] M. Brandstötter, T. Komenda, F. Ranz, P. Wedenig, H. Gattringer, L. Kaiser, G. Breitenhuber, A. Schlotzhauer, A. Müller, and M. Hofbaur, "Versatile collaborative robot applications through safety-rated modification limits," in *Proc. Int. Conf. Robot. Alpe-Adria Danube Region (RAAD)*, Kaiserslautern, Germany, 2019, pp. 438–446.

[293] N. Duzmanska, P. Strojny, and A. Strojny, "Can simulator sickness be avoided? A review on temporal aspects of simulator sickness," *Frontiers Psychol.*, vol. 9, p. 2132, Nov. 2018.

[294] D. Saredakis, A. Szpak, B. Birckhead, H. A. D. Keage, A. Rizzo, and T. Loetscher, "Factors associated with virtual reality sickness in head-mounted displays: A systematic review and meta-analysis," *Frontiers Hum. Neurosci.*, vol. 14, p. 96, Mar. 2020.

[295] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, 1993.

[296] (Dec. 2021). *Benchmark, N. and ADJ*. Oxford University Press. Accessed: Feb. 15, 2022. [Online]. Available: https://www.oed.com/view/Entry/17612

[297] Y. Sun, J. Falco, N. Cheng, H. R. Choi, E. D. Engeberg, N. Pollard, M. Roa, and Z. Xia, "Robotic grasping and manipulation competition: Task pool," *Commun. Comput. Inf. Sci.*, vol. 816, pp. 1–18, Jan. 2018.

[298] H. Mnyusiwalla, P. Triantafyllou, P. Sotiropoulos, M. A. Roa, W. Friedl, A. M. Sundaram, D. Russell, and G. Deacon, "A bin-picking benchmark for systematic evaluation of robotic pick-and-place systems," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1389–1396, Apr. 2020.

[299] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first Amazon picking challenge," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, Jan. 2018.

[300] P. So, J. Wittmann, P. Ruhkamp, A. Sarabakha, and S. Haddadin, "Towards remote robotic competitions: An internet-connected task board and dashboard," 2022, *arXiv:2201.09565*.

[301] K. Collins, A. J. Palmer, and K. Rathmill, "The development of a European benchmark for the comparison of assembly robot programming systems," in *Robot Technology and Applications*. Berlin, Germany: Springer, 1985, pp. 187–199.

[302] S. Zeylikman, S. Widder, A. Roncone, O. Mangin, and B. Scassellati, "The HRC model set for human–robot collaboration research," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1845–1852.

[303] G. Sarthou, A. Mayima, G. Buisan, K. Belhassein, and A. Clodic, "The director task: A psychology-inspired task to assess cognitive and interactive robot architectures," in *Proc. 30th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Vancouver, BC, Canada, Aug. 2021, pp. 770–777.

[304] F. Formica, S. Vaghi, N. Lucci, and A. M. Zanchettin, "Neural networks based human intent prediction for collaborative robotics applications," in *Proc. 20th Int. Conf. Adv. Robot. (ICAR)*, Ljubljana: Institute of Electrical and Electronics Engineers Inc., Dec. 2021, pp. 1018–1023.

[305] C. Chao and A. Thomaz, "Timed Petri nets for fluent turn-taking over multimodal interaction resources in human–robot collaboration," *Int. J. Robot. Res.*, vol. 35, no. 11, pp. 1330–1353, Sep. 2016.

[306] M. E. Foster, N. Giuliani, A. Isard, C. Matheson, J. Oberlander, and A. Knoll, "Evaluating description and reference strategies in a cooperative human–robot dialogue system," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Pasadena, CA, USA, 2009, pp. 1–6.

[307] L. Gualtieri, E. Rauch, and R. Vidoni, "Development and validation of guidelines for safety in human–robot collaborative assembly systems," *Comput. Ind. Eng.*, vol. 163, Jan. 2022, Art. no. 107801.

[308] A. Roncone, O. Mangin, and B. Scassellati, "Transparent role assignment and task allocation in human robot collaboration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Singapore, May 2017, pp. 1014–1021.

[309] T. Hamabe, H. Goto, and J. Miura, "A programming by demonstration system for human–robot collaborative assembly tasks," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Zhuhai, Dec. 2015, pp. 1195–1201.

[310] A. Pupa, W. Van Dijk, C. Brekelmans, and C. Secchi, "A resilient and effective task scheduling approach for industrial human–robot collaboration," *Sensors*, vol. 22, no. 13, p. 4901, Jun. 2022.

[311] M. Cramer, K. Kellens, and E. Demeester, "Probabilistic decision model for adaptive task planning in human–robot collaborative assembly based on designer and operator intents," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7325–7332, Oct. 2021.

[312] A. Angleraud, R. Codd-Downey, M. Netzev, Q. Houbre, and R. Pieters, "Virtual teaching for assembly tasks planning," in *Proc. IEEE Int. Conf. Hum.-Mach. Syst. (ICHMS)*, Sep. 2020, pp. 1–8.

[313] A. Rajavenkatanarayanan, H. R. Nambiappan, M. Kyrarini, and F. Makedon, "Towards a real-time cognitive load assessment system for industrial human–robot cooperation," in *Proc. 29th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Naples, Italy, Aug. 2020, pp. 698–705.
[314] V. Havard, B. Jeanne, M. Lacomblez, and D. Baudry, "Digital twin and virtual reality: A co-simulation environment for design and assessment of industrial workstations," *Prod. Manuf. Res.*, vol. 7, no. 1, pp. 472–489, 2019.
[315] K. N. Kaipa, C. W. Morato, and S. K. Gupta, "Design of hybrid cells to facilitate safe and efficient human–robot collaboration during assembly operations," *J. Comput. Inf. Sci. Eng.*, vol. 18, no. 3, Jun. 2018, Art. no. 031004.
[316] R. S. Wilcox, S. Nikolaidis, and J. A. Shah, "Optimization of temporal dynamics for adaptive human–robot interaction in assembly manufacturing," in *Robotics: Science and Systems*. Cambridge, MA, USA: MIT Press, 2012.
[317] H. Liu and L. Wang, "An AR-based worker support system for human–robot collaboration," *Proc. Manuf.*, vol. 11, pp. 22–30, Jan. 2017.
[318] R. Maderna, M. Pozzi, A. M. Zanchettin, P. Rocco, and D. Prattichizzo, "Flexible scheduling and tactile communication for humanbrobot collaboration," *Robot. Comput.-Integr. Manuf.*, vol. 73, Feb. 2022, Art. no. 102233.
[319] D. Kofer, C. Bergner, C. Deuerlein, R. Schmidt-Vollus, and P. Heß, "Human–robot-collaboration: Innovative processes, from research to series standard," *Proc. CIRP*, vol. 97, pp. 98–103, Jan. 2021.
[320] A. Casalino, A. M. Zanchettin, L. Piroddi, and P. Rocco, "Optimal scheduling of human–robot collaborative assembly operations with time Petri nets," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 1, pp. 70–84, Jan. 2021.
[321] E. Martin, D. R. Lyon, and B. T. Schreiber, "Designing synthetic tasks for human factors research: An application to uninhabited air vehicles," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 1, Nov. 1998, pp. 123–127.
[322] L. R. Elliott, M. Dalrymple, J. W. Regian, and S. G. Schiflett, "Scaling scenarios for synthetic task environments: Issues related to fidelity and validity," in *Proc. 45th Annu. Meeting Human Factors Ergonom. Soc.*, 2001, pp. 377–381.
[323] N. J. Cooke and S. M. Shope, "Designing a synthetic task environment," in *Scaled Worlds: Development, Validation, and Application*, S. G. Schiflett, L. R. Elliott, E. Salas, and M. D. Coovert, Eds. Surrey, U.K.: Ashgate, 2004, pp. 263–278.
[324] C. Lenz, "Context-aware human–robot collaboration as a basis for future cognitive factories," Ph.D. dissertation, Dept. Comput. Eng., Tech. Univ. Munich, München, Germany, 2011.

DOMINIK RIEDELBAUCH received the master's and Ph.D. degrees in computer science from the University of Bayreuth, Germany, in 2016 and 2022, respectively. He was a Postdoctoral Researcher with the Chair for Applied Computer Science III (Robotics and Embedded Systems), from 2020 to 2022. His research interests include dynamic human-robot teaming, cognitive human models, and simulation-based and replicable experiments in human–robot interaction research.

NICO HÖLLERICH received the master's degree in computer science from the University of Bayreuth, Germany, in 2018, where he is currently pursuing the Ph.D. degree with the Chair for Applied Computer Science III (Robotics and Embedded Systems). His current research interests include dynamic human–robot teaming enhanced by artificial intelligence, replicable experiments, and workspace observance under uncertainties.

DOMINIK HENRICH received the Diploma degree in computer science from the University of Karlsruhe, Germany, in 1991, and the Ph.D. degree. From 1992 to 1994, he was supported by the German Research Foundation scholarship. From 1996 to 1999, he built up the "Parallel Processing and Robotics" Research Group, Institute for Process Control. In 1998, he received a STA Fellowship with the Electrotechnical Laboratory of the Ministry of International Trade and Industry (MITI), Japan. From 1999 to 2003, he was a Professor of the "Embedded Systems and Robotics" Group, University of Kaiserslautern. He has been holding the Chair for Applied Computer Science III (Robotics and Embedded Systems), University of Bayreuth, Germany, since August 2003. His research interests include collision detection, motion planning, room surveillance, self-adapting robots, sensor-based manipulation, intuitive robot programming, and human/robot cooperation.