Towards Effective Interface Designs for Collaborative HRI in Manufacturing: Metrics and Measures

JEREMY A. MARVEL, SHELLY BAGCHI, MEGAN ZIMMERMAN, and
BRIAN ANTONISHEK, U.S. National Institute of Standards and Technology, USA

We present a comprehensive framework and test methodology for the evaluation of human-machine interfaces (HMI) and human-robot interactions (HRI) in collaborative manufacturing applications. An overview of the challenges that face current- and next-generation collaborative robot systems is presented, specifically focused on the interactions between man and machine, and a series of objectively quantitative and subjectively qualitative metrics are given to guide the development and assessment of interfaces and interactions. A generalized set of guidelines for the design of HMI is also proposed to address these challenges and thereby enable effective and intuitive diagnostics and error corrections when process failures occur. These guidelines are aimed at aiding researchers in developing effective interface and interaction technologies, maximizing operator situation awareness in human-robot collaborative manufacturing teams, promoting effective process and system diagnostics reporting, and enabling faster responses to equipment or application errors.

CCS Concepts: • **Human-centered computing** → **Collaborative interaction**; **User interface design**; **User centered design**; • **Computer systems organization** → **External interfaces for robotics**;

Additional Key Words and Phrases: Test methods and metrics, repeatability studies, peformance measures, benchmarking, use cases, data sets

ACM Reference format:
Jeremy A. Marvel, Shelly Bagchi, Megan Zimmerman, and Brian Antonishek. 2020. Towards Effective Interface Designs for Collaborative HRI in Manufacturing: Metrics and Measures. *Trans. Hum.-Robot Interact.* 9, 4, Article 25 (May 2020), 55 pages.
https://doi.org/10.1145/3385009

1 INTRODUCTION

The use of robotics in U.S. manufacturing is on the rise [Doyle 2015; Orcutt 2014] and is expected to continue growing exponentially through the coming years [Wood 2014]. Although the number of robots in manufacturing has been increasing steadily over several decades, the current growth trend is notable given the number and variety of industries using robots. While the large-scale automotive and electronics industries continue to dominate the market on robotics use, the use of robotics in smaller-scale applications is slowly rising [Marvel et al. 2015b]. Collaborative manufacturing robots—relative newcomers to the robotics market—are expected to constitute a considerable

Authors' addresses: J. A. Marvel, S. Bagchi, M. Zimmerman, and B. Antonishek, U.S. National Institute of Standards and Technology, 100 Bureau Dr., Stop 8230, Gaithersburg, MD, 20899; emails: {jeremy.marvel, shelly.bagchi, megan.zimmerman, brian.antonishek}@nist.gov.

2573-9522/2020/05-ART25
https://doi.org/10.1145/3385009

percentage of this population growth, particularly among small- and medium-sized manufacturers (SMMs), due to the cost benefits associated with the perceived inherent safety [Marvel 2014a].

Currently, the "collaborative" nature of collaborative robots is one merely of proximity. These robot platforms are used in traditional automation applications and are largely operating independently of the environments and events surrounding them. While this may suffice for contemporary applications, it limits the potential of the technology. To be more impactful in SMM applications, robots must assume roles of intelligent, assistive tools for skilled human labor. This necessitates interactions beyond mere collision avoidance and touch sensitivity and will require advances in system and process diagnostics, prognostics, and situation awareness (SA). Moreover, these robots must possess the tools to effectively communicate all of this information to promote task efficiency, fault avoidance, and error recovery.

To transition from the traditional paradigm of manufacturing robotics to a new generation of user-friendly designs, two criteria must be met. First, it must be shown that the user-friendly designs are actually more effective or efficient. While more advanced robotic functions may be less oriented toward casual users, basic programming, maintenance, and production processes are expected to benefit in terms of learning time, time to completion, and quality of results. Second, it must be established that the operators' responses to the technology are positive. Effectiveness, efficiency, and user satisfaction form the basis for human-centered design principles, as described in ISO 9241-210:2010 [ISO 2010], which outlines the specifications and recommendations for the development of computer-based interactive systems (e.g., software, websites, mobile phones, and vending machines). As a general rule, people will not move toward something new if they are not attracted or inspired by it.

This report proposes a metrology-based framework for developing effective mechanisms for information exchange and collaboration in human-robot teams. This effort is motivated by the need for a common basis for evaluating and comparing human-robot interactions and interfaces, which thus enables the advancement of collaborative robot technologies. There exists a rift between the cutting edge in HRI research and the current state of practice within industry. This article aims to bridge this gap and provide researchers with the insights of industry's needs when designing novel HRI technologies, and to arm industry with the tools and techniques used in emerging systems.

It is important to draw a distinction between the interface and the interaction a user has when working with a given system. Terms such as "human-machine interface" (HMI), "human-computer interface" (HCI), or, more generically, "user interface" (UI) are used to describe the mechanisms through which humans and machines communicate. Such interfaces can consist of monitors, lights, keyboards, switches, joysticks, or any variety of sensors that share inputs and information between the human user and the machine and its software. In contrast, the "user experience" (UX), or the "human-robot interaction" (HRI) specific to the domain of robotics, is a means of describing the nature and quality of the information being shared in addition to its impacts on both the user and the system.

While this report highlights lessons learned across multiple domains of HRI, it aims to provide guidelines specifically targeted at the manufacturing domain. Therefore, for the purposes of this report, we use the standardized robotics vocabulary defined in ISO 8373:2012 [ISO 2012]. Section 2 outlines the communication challenges faced by current and next generation collaborative robot systems. Section 3 discusses HRI in the manufacturing context. Section 5.3 outlines the metrics and bases for effective system and process diagnostics communications in HRI, while Section 5 provides a generalized framework to guide the development of HMI design for next-generation collaborative robot systems. Section 6 concludes the article with closing observations and general comments.

Throughout this report, a number of sources for metrics, test methods, and rubrics are cited to provide bases for the recommendations made in Section 5. In many cases, recent studies and

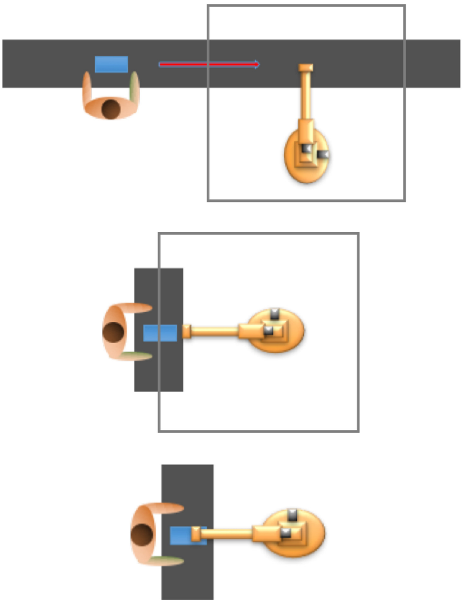

Fig. 1. In sequential collaborations (top) humans and robots are in a shared process flow of a workpiece, but are not necessarily co-located. Simultaneous collaborations (middle) see humans and robots working on different parts of the workpiece at the same time. Humans and robots in supportive collaborations (bottom) must work together to accomplish shared task goals.

surveys have used as their bases of evaluation the metrics established in older texts. In the process of writing this report, it was decided that citations of the original sources of these metrics was preferable to using the most recent instances of the metrics' use. Though the newer studies provide insightful instances of new interfaces, algorithms, and systems, the focus of this article is ultimately on the test methods and metrics. As such, only the most relevant sources are cited.

2 COLLABORATIVE ROBOTICS IN MANUFACTURING

At the center of the discussion concerning effective HRI is the notion that there must be some level of interaction between a human and a robot. This implies some level of collaborative functionality between a human and a machine. There are effectively four degrees of human-robot collaboration [Helms et al. 2002], three of which are illustrated in Figure 1:

(1) Separate: The human and robot tasks are kept apart; they do not share workspaces, tools, or work pieces.
(2) Sequential: The human and robot tasks are consecutively completed. The workspaces, tools, and work pieces may be shared, but there is a strict serialization of the tasks such that any sharing is temporally separated.
(3) Simultaneous: The human and robot tasks are executed concurrently, and may involve working on different parts of the same work piece, but are focused on achieving separate task goals.
(4) Supportive: The human and the robot work together at the same time and with the same work piece to complete a common task.

Historically, human-robot collaboration has been relegated to "separate" applications. The current state-of-the-art in terms of human-robot teams may be classified as "sequential," given that automation and human labor are frequently seen as successive links in the production chain. With the recent introduction of collaborative industrial robots to the market, simultaneous collaborations are now feasible, albeit not common. Future generations of collaborative robot systems are expected to take "supportive" roles, but such functionality is beyond the capabilities of current technology.

Enabling supportive collaborations requires advances in environmental and process sensing, validated safety, SA models, and intuitive interfaces. Integrating robots into human teams will provide unique challenges stemming from the anticipated lack of robotics experience or expertise on the part of the human operators, issues regarding ease of use or programming, and general distrust of robots and their capabilities [Groom and Nass 2007]. Interacting with a collaborative robot must be as safe, natural, and effective as interactions with another human.

Robots currently marketed as being "collaborative" are designed specifically to be safe to operate around human operators. Their requirements, defined in ISO/TS 15066:2016 [ISO 2016] and discussed in more detail in Section 3.3, are functionally disconnected from the degrees of collaboration described above. To this end, it is worth pointing out the distinction between the terms "collaborative robot" and "collaborative operation" from ISO 8373:2012. A collaborative robot is a "robot designed for direct interaction with a human," while a collaborative operation is a "state in which *purposely designed* robots work in direct cooperation with a human within a defined workspace" (emphasis added). Although current trends indicate a steady increase in collaborative robot sales over the next several years, few of these robots will be integrated into collaborative operations. Instead, these robots are likely to be installed as stopgap automation solutions in human-occupied environments, specifically because they are safer and generally easier to use than their more traditional robot brethren [Marvel 2014a].

2.1 Robotic Diagnostics

The first-level interactions between operators and the robotics revolve around diagnostics of robots and robotic processes. This defines the level of most human-robot interactions in industrial settings. With the aforementioned forecast in robotics growth, the challenge of integrating collaborative robots into historically manual processes will drive the development of new technologies that enable more seamless and efficient hardware interoperability. With the steady adoption of just-in-time manufacturing processes and Internet of Things [Atzori et al. 2010] integration, however, the increased interconnectivity of machine tools, robots, and automation systems presents new potential failure points and new challenges in terms of guaranteeing performance and minimizing downtime due to said failures. Failures must be diagnosed quickly, and the necessary steps to correct for failures must be presented to ensure minimal loss of productivity.

With constant health monitoring of hardware and process performance, performance degradation may be identified early and failures avoided altogether. However, the requisite system and process performance data must be presented in clear and meaningful ways for such information to be useful. As the complexity of integrated systems increases, providing meaningful feedback to operators in the event of system or process failures becomes increasingly challenging.

As with any machine tool, robotic systems are integrated with fault identification and diagnostic tools for the hardware and software. Different mechanisms have been proposed for capturing and evaluating robot system health, ranging from basic state representations [Heddy et al. 2015] to observer-based [Caccavale and Walker 1997] and model-based diagnostic tools [Liu and Coghill 2005]. Most monitor basic key performance indicators such as robot dynamics, kinematics, joint torques, joint temperatures, cycle times, and integrity of redundant communication, sensor, and safety channels.

On the whole, hardware failures of robotic systems are a rare occurrence. Over a sufficiently long period of time, brakes may fail, encoders degrade, and the accuracy and repeatability of the robot as a whole decline. However, such degradation occurs over years and is typically preceded by a steady reduction in efficiency. Short-term process failures, in contrast, are driven by damage to tooling and by logic errors caused by sensing and software malfunctions. Such errors are not inherently detectable by the robot's controller, but may be inferred from deviations from nominal performances (e.g., increasing times to perform assemblies, upward trends in the number of cycle failures, or significant deviations of incurred forces, torques, or joint currents). When these failures occur it is paramount that the root causes are identified and corrected quickly, efficiently, and correctly.

2.2 Human Roles in Modern Robotic Collaborations

In collaborative human-robot teams, the roles of the human operator are typically limited to supervision/operation, programming, and integration/maintenance. All of these are largely to be considered *separate* collaborations. Programming, for instance, is a front-loaded effort. Once a robot begins production, the programming element is concluded until the robots are retasked, or an error arises and the programmer is required to return to find and fix a logic error.

Similarly, integration and maintenance are one-off interactions performed as needed. Integration occurs before programming begins, and maintenance is overwhelmingly preventative or reactive in nature. Preventative maintenance entails doing routine actions to prevent common faults by lubricating joints, periodic cleaning/replacing components and tools, and so on. Reactive maintenance occurs when something breaks and may either be done *in situ* with the workcell taken out of production, or the equipment is replaced with a spare and the faulty equipment is taken offline for refurbishing.

Supervision and operation entails observing the state of the equipment and processes and stepping in periodically for inspection, task assignments, or process maintenance (e.g., clearing a workcell of debris). This is a perpetual level of involved interaction. As such, the role of the operator is that which is typically associated with HRI, whereas the programming, maintenance, and integration roles are impacted more by HMI. Each of the three roles requires different degrees and content of real-time feedback. In the next section, we will discuss the interfaces and natures of interactions between humans and robots. The nature of HRI is expected to fundamentally change as robot technology evolves, and the current characterization of human roles may someday be an outdated archetype. Providing the evaluative tools for optimizing HRI and HMI will drive innovation, and it is expected the metrics discussed in this survey will also evolve as a result.

With the latest generation of industrial robots with collaborative safety functions, hand-guiding an active robot can be a safe and viable solution for human-robot teaming. In some modern factory environments, the use of collaborative robots is focused on assisting the human operator for material handling operations [Gambao et al. 2012; Michalos et al. 2018; Shi et al. 2012]. In such applications, the robot acts as an assisted lift device or dynamic fixture, enabling operations such as collaborative inspection, manipulation, assembly, and finishing.

Recent studies indicate that dramatic shifts from purely manual processes to mixed manual-automated tasks may have a negative impact on the human workforce. Specifically, the human workers may express performance impacts, including speed and process flexibility, as a result of having to work with robots [Weiss and Huber 2016]. To minimize the potential negative impacts of integrating human-robot teams, it has been found that maintaining the "human" aspects of social interaction, autonomy, and problem solving are important for operator buy-in and positive user experience [Welfare et al. 2019].

3 HRI IN MANUFACTURING

Robotic systems for manufacturing have experienced an interesting evolution in terms of applications since their introduction over half a century ago, while the underlying technologies have largely remained unchanged. Although the capabilities, controllers, and interfaces for robot systems have changed significantly, the industrial robot model has not. Robots remain collections of actuated drive units (pneumatic, hydraulic, or electric) connected via rigid links to move specialized tooling around a large work volume.

Despite this lack of advancement, the idea that a robot is a standalone entity that operates independently within the environment is an outdated archetype. With agile manufacturing, evolving paradigms of "smart" factory environments, and the advancement of Internet of Things into manufacturing methods, the tools, systems, and processes have become increasingly interconnected and interdependent.

To develop the requisite test methods and metrics, it is first necessary to establish what effective communications actually entail and to identify the interfaces through which humans and robots interact. Beginning with first principles, one must identify (1) the intended interactions between a human and a robot and (2) the purpose of the information exchanged between humans and robots. Both elements are largely determined by the application domain and the roles of the individual people and robots [Driewer et al. 2007] and must be adapted to different contexts. While much of the published literature on effective HRI focuses primarily on social, teleoperated, and theoretical contexts, key performance indicators in manufacturing are driven from the perspective of productivity and throughput [Kangru et al. 2018; Thomas 2018; Weiss et al. 2019, 2013, 2018; Zhu et al. 2018].

3.1 Human-machine Interfaces

Industrial robot controllers typically come with teach pendants (hand-held control boxes developed by the robot manufacturer) that serve as the principal interface for calibrating, moving, and programming the robots. Some modern collaborative robot designs forego the teach pendant entirely and rely on advanced sensing, joint compliance, and artificial intelligence to enable human-guided task learning. Ancillary equipment such as light towers and sirens may be connected to the robot to give visual and audible cues of robot state. In workcells where the robot is the only complex machine, the robot's controller may also double as the controller for connected tools and feeder equipment (e.g., conveyors and loaders). However, if integrated into a larger system of interconnected, controlled tools, the controller may be configured to be slave to a master programmable logic control (PLC) used for controlling process flow and maintaining workplace safety.

For maintenance and programming functions, the primary interface is the robot's teach pendant, which provides direct access to the robot's logic. Content provision may be complicated, however, in multi-robot workcells consisting of heterogeneous robot systems (i.e., workcells consisting of robot systems from different manufacturers). For multi-robot workcells consisting of a single manufacturer's products, a singular interface is learned for system integration, programming, and feedback. Heterogeneous robot configurations pose unique challenges due to the lack of vendor support for integrating dissimilar operating systems and control interfaces. Providing for such robot-robot collaborations will be necessary for future flexible manufacturing solutions. Efforts to expand standards for machine tool feedback mechanisms (e.g., MTConnect [Vijayaraghavan et al. 2008] and EtherNet/IP [Brooks 2001]) to be fully inclusive of robot systems, or for developing common command interfaces (e.g., LabVIEW for Industrial Robotics [National Instruments 2016] or the Industrial build of the Robot Operating System [Edwards and Lewis 2012]) are anticipated to be highly impactful as enabling technologies for easing integration burdens.

In contrast, operational feedback is more likely to come via an integrated PLC, which fuses information from multiple data sources. While the PLC may offer some functional control for process optimization, it tends to be a limited interface for collaborative operations. Thus, the roles of the operator are usually relegated to sequential or simultaneous collaborations.

Alternative interfaces and tools are required to move beyond simple coexistence and toward supportive collaborative roles where the robots will be intelligent, assistive tools for operators. What is needed is an interface specifically designed to increase efficiency and productivity for both the robots and the operators. Such a tool would enable a single, skilled worker capable of performing better and more productively than they would be unassisted. Such an interface would increase basic production capabilities, drive in-line quality control, and provide real-time feedback and situation awareness for improved prognostics and health monitoring of equipment and processes. However, it is not the operator's needs, but rather cost reduction, that drives many HMI design decisions [Shneiderman 1998]. As such, the recommendations for stakeholder-influenced, iterative interface design processes from ISO 9241-210 are not always heeded when the HMI is being built.

3.2 Interactions in Human-robot Teams

Most of the available literature regarding effective HRI is based on field studies involving teleoperated robot applications such as urban search and rescue (USAR) and medical operations. Understanding the fundamental differences between teleoperated robot applications and collaborative manufacturing applications is key to digesting and making HRI recommendations based on such field studies.

Within the realm of manufacturing applications, it is assumed that robots and humans are colocated, which naturally provides contextual clues not present in teleoperated human-robot teams. In a study by Murphy and Burke [2005], for instance, it was determined that over 60% of the remote operator's time is spent establishing and maintaining SA. The authors go on to state that human-human teams are nearly an order of magnitude more effective at accomplishing complex tasks than human-robot teams due largely to issues with establishing SA.

In search and rescue applications, the role of robots is largely reduced to being a source of information for task operators, while physical work is a relatively lower priority. This information gathering is impacted by several factors, including time delays, sensing and display limitations, training, and restrictions on information bandwidth.

3.3 Collaborative Robot Safety

Any interactions between operators and robots are expected to be safe. Current safety guidelines, defined internationally in ISO 10218-1:2011 [ISO 2011b] for robots, and ISO 10218-2:2011 [ISO 2011a] for robot systems, require traditional industrial robots remain separate from operators. However, collaborative robots, as mentioned earlier, are designed with direct interaction in mind. The safety features of these robots are dictated by ISO/TS 15066:2016 [ISO 2016], which defines four collaborative safety modes:

- Safety-rated monitored stop, which is effectively a software-based "pause" function for collaborative robot programs.
- Hand-guiding, which enables an operator to directly steer the motions of the robot while in automatic mode.
- Speed and separation monitoring (SSM), which seeks to keep a safe separation distance between the operator and an active robot.
- Power and force limiting (PFL), which seeks to ensure operator safety in circumstances where the operator and active robot make physical contact.

There have been efforts to capture, characterize, and report the operational functionality of these safety requirements (e.g., Marvel [2013]; Marvel and Norcross [2017]; Zanchettin et al. [2016] for SSM, and [Falco et al. 2012; Matthias et al. 2014] for PFL). Many such efforts, however, are limited by the technological shortcomings of sensor systems designed to detect, identify, and model humans [Marvel and Norcross 2017]. From a metrology perspective, the assessment of SSM is contingent on the separation distance, d_{sep}, and speed of approach, v_r, between the closest part of the human and the closest part of the active manipulator.

In contrast, the measurement of PFL is contingent on the type of impact occurring between the robot and the human. Impacts are classified as either transient or quasi-static, depending on the conditions of the impact. Quasi-static cases involve part(s) of the body being compressed by the robot against a rigid surface for a period $\geq$0.5 s. Transient impacts are short duration (<0.5 s) in which either the robot or the human can recoil from the contact. In both cases, force (N) and pressure (N/cm^2) are measured at the point of contact (both of the human and of the robot), but are compared against different limits depending on the body region. Force and pressure limits are lower for quasi-static impacts than those of transient impacts. Provided these values are lower than the limits established in ISO [2016], the impact event is considered "safe."

3.4 A Basis for More Effective HRI

The modern workforce is becoming increasingly technologically adept, and, with this, more discerning about the technologies with which they interact on a daily basis. In this information age, it is expected that relevant information will be presented in real-time and in formats that make it easy to digest said information quickly. In many ways, manufacturing technology has failed to evolve along these lines, and it is not uncommon for generational differences to result in a sense of malaise when working with the machinery. Most manufacturing machinery is focused on operations, whereas the current trend in consumer equipment is human-centricity. Rather than emphasizing human awareness by keeping operators informed, trained, or "in-the-loop," or maintaining application or automation flexibility [Wickens et al. 2004], the traditional performance-based archetypes of robot repeatability and accuracy still drive HMI designs.

It is expected that much of the indirect interactions between the operator and robot will be accomplished via some form of HMI, such as a computer monitor, teach pendant, or other visible indicators such as light towers. Direct interactions will be based on physical interactions with the HMI or the robots themselves. At no point should SA be compromised, however, and the HRI is expected to provide contextual information regarding task and robot state (i.e., system diagnostics). In general, HRI should support (or at worst be agnostic to the support) of human-robot, human-human, robot-human (i.e., the robot's awareness of the human), and robot-robot awareness [Drury et al. 2003]. This awareness is accommodated by shared visual information that is used to accommodate the following activities:

- Build shared mental models and facilitate team coordination;
- Coordinate team activities through targeted and non-targeted (i.e., generalized) communication;
- Increase the efficiency of team communication; and
- Perform the task.

These activities are enabled in effective HRI via well-designed HMI and intelligent software. People are better aware of problems with a process if they are actively participating in that process, as opposed to casual observers or maintenance crews coming in after an error has already occurred [Wickens et al. 2004]. This phenomenon, known as the generation effect, impacts the SA of operators who are not actively involved with a robotic process. HRI must therefore accommodate

Table 1. Response Robotics HRI Metrics [Steinfeld et al. 2006]

Subject	Category	Metrics
Team	Quantitative performance	Effectiveness
		Efficiency
	Subjective rating	Effort quality
	Utility of mixed initiative	% Robot assistance requests
		% Human assistance requests
		Number non-critical human interruptions
		Asset utilization of "functional primitives" [Rodriguez and Weisbin 2003]
		Interactive effort [Olsen and Goodrich 2003]
Operator	Situation awareness	Situation awareness global assessment technique [Endsley 1995a]
	Workload	NASA-Task Load index (NASA-TLX, [Hart and Staveland 1988])
		Physiological measures
	Mental model accuracy	Conceptual
		Movement
		Spatial
		Modality
Robot	Self awareness	Intrinsic limitations
		Self-monitoring and modeling
		Fault detection, isolation, recovery
	Human awareness	Human-oriented perception
		Human modeling and monitoring
		Human sensitivity
	Autonomy	Neglect tolerance [Olsen and Goodrich 2003]

operator engagement if it is expected the operator will be an effective, contributing member in the human-robot team.

4 PERFORMANCE METRICS

To determine whether or not a given robot interface is effective, one must first have metrics by which said interface may be assessed. In this section, we categorically highlight many of the metrics and test methods used to assess and assure the effective application of robots to collaborative tasks.

Efforts to assemble metrics for HRI often focus on a specific task domain. For example, Steinfeld et al. [2006] compiled a collection of common metrics for evaluating HRI for response robotics. In this model, the HRI metrics are separated into three broad categories: team performance, operator performance, and robot performance (Table 1). While these metrics are specifically targeting systems consisting of a single operator and robot, many are scalable to multi-robot, multi-human scenarios, and all are applicable to manufacturing robotics. Such collections of metrics are fairly common in the field of HRI and touch on multiple sub-fields of research that will be discussed in detail in the coming sections.

While many of these metrics are subjectively qualifiable, the end goal is to also provide objectively quantifiable measures that can be used to evaluate collaborative robot HRI without relying on post-process surveys and assessments. The reasons for this are two-fold. First, subjective measures, while useful for capturing the operator's perspectives of robot interactions, require feedback that is difficult to decompose and generalize. People are not necessarily the best judges of their own performance, and repeatability of results may be compromised by external factors not captured in surveys (including the validity of the survey itself). Objective measures, while not necessarily capable of modeling emotions, stress, or effort, are also less likely to be influenced by such factors. Second, such assessments take time out of an otherwise busy work day to complete. Given also that most subjectively qualitative measures may be recorded only as post-factor assessments, they may also suffer from the participants simply wanting to leave. As such, there is likely to be pressure to promote timeliness over thoroughness, which may negatively compromise the final results of evaluations. Objectively quantitative measures, in contrast, may be captured automatically and in real-time. Ultimately, both quantitative and qualitative measures are necessary, and the goal of this report is to provide both whenever feasible.

4.1 The Use of Surveys in HRI Research

It is common for HRI studies to use some form of survey or questionnaire to capture user preferences. Frequently, these surveys are custom-designed for the specific study being conducted. As such, the surveys—and typically the subsequent results—are not broadly applicable across studies. Occasionally, however, surveys are specifically designed to be used as repeatable instruments. These surveys are intended to capture specific elements of the HRI inquiry and may be combined with additional surveys to provide insights across multiple metrics.

For example, an analysis of 125 papers and extended abstracts from the joint 2019 Association of Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE) International Conference of Human-Robot Interaction provided an insight into the current trends of survey utilization. Of these papers, 84.8% used subjective surveys to capture participant preferences. Of these, 94.8% used surveys customized for the respective studies. This reliance on customized surveys by such a significant percentage of HRI research can make it difficult to extrapolate meaningful metrics for generic HRI use cases. To this end, the focus of this article is largely limited to the survey instruments that are intended for broad applicability.

The use of pre-established survey instruments from the 2019 ACM/IEEE HRI Conference accounted for 19.0% of all subjective measures. In all, there were 23 such instruments used for a total of 43 citations. Of these instruments, however, only 5 were used more than once, meaning 41.9% of all cited survey instruments were used only by a single paper. This is in contrast with the 2015 ACM/IEEE HRI conference proceedings, in which only 19 of the 165 papers (11.5%) cited established survey instruments. Over these 19 papers, there were 33 citations of instruments used, of which only 3 surveys were used by more than 1 paper. As such, a little more than 84% of all cited survey instruments in 2015 were used only by a single paper.

Some of these single-use citations from the 2019 proceedings were specific for certain domains such as individual feelings toward nature [Shibata 2016] or assessments of specific groups of people (e.g., elderly care [Wang 2005]). It is worth noting that many of these survey instruments include measurements of similar concepts or reactions to technology. For example, Table 2 provides an overview of the metrics used by a number of the cited survey instruments from the 2019 ACM/IEEE HRI Conference. While a number of metrics are covered by a single survey, many are represented across multiple surveys.

While some of these surveys are quite dated, many of these older evaluation tools continue to be used due to two principal factors: (1) they have been validated, and their strengths and weaknesses are known, and (2) the aspects that they attempt to capture are still highly relevant to modern HRI. Some recent studies have attempted to augment existing surveys with additional questionnaires, but these augmentations have a tendency to be highly application-specific.

Table 2. Subsample of Subjective Factors Captured in Cited Survey Instruments

Measure	Survey(s)
Adaptiveness	Almere Model [Heerink et al. 2010]
Animacy	Godspeed [Bartneck et al. 2009]
Anxiety	Almere Model; Fear of Negative Evaluation Scale [Ishikawa et al. 1992]; Multidimensional Robot Attitude Scale [Ninomiya et al. 2015]; Negative Attitude toward Robots Scale (NARS) [Nomura et al. 2006]; Social Avoidance and Distress Scale (SADS) [Ishikawa et al. 1992]
Arousal	Self Assessment Manikin Instrument [Bradley and Lang 1994]
Anthropomorphism	BEHAVE-II [Joosse et al. 2013]; Godspeed; Multidimensional Robot Attitude Scale
Attitude Towards Technology	Almere Model; BEHAVE-II
Attractiveness/Appearance	BEHAVE-II; Multidimensional Robot Attitude Scale
Believability	Multi-Dimensional Measure of Trust (MDMT) [Ullman and Malle 2018]
Competence	Almere Model; Robot Social Attributes Scale (RoSAS) [Carpinella et al. 2017]; MDMT
Complexity	System Usability Scale [Brooke 1996]
Confidence in Use	System Usability Scale; MDMT
Control/Choice	Multidimensional Robot Attitude Scale; NARS
Cumbersome	System Usability Scale
Discomfort	RoSAS; NARS; SADS
Dominance	Self Assessment Manikin Instrument
Ease of Learning	System Usability Scale; Multidimensional Robot Attitude Scale
Ease of Use	System Usability Scale; Almere Model; Multidimensional Robot Attitude Scale
Effort	NASA TLX; Multidimensional Robot Attitude Scale
Environmental Facilitation	Almere Model; Multidimensional Robot Attitude Scale; SADS
Ethics	MDMT; Negative Attitude toward Robots Scale
Familiarity	Multidimensional Robot Attitude Scale; NARS
Frustration	NASA TLX
Inconsistency	System Usability Scale
Intelligence	Godspeed
Intent/Interest in Using	Almere Model; Multidimensional Robot Attitude Scale; System Usability Scale;
Learning Required	Multidimensional Robot Attitude Scale; System Usability Scale
Likability	Almere Model; BEHAVE-II; Godspeed; NARS
Mental Demand	NASA TLX; SADS
Need Help Using	System Usability Scale
Performance	NASA TLX
Physical Demand	NASA TLX; Simulator Sickness Questionnaire [Kennedy et al. 1993]
Pleasure	NARS; Self Assessment Manikin Instrument; SADS
Presence	NARS; SADS
Quality of Integration	System Usability Scale
Safety	Godspeed
Sociability/Warmth	Almere Model; RoSAS; NARS; SADS
Social Influence	Almere Model; Multidimensional Robot Attitude Scale; SADS
Temporal Demand	NASA TLX
Trust	Almere Model; BEHAVE-II; MDMT; NARS
Usefulness	Multidimensional Robot Attitude Scale
Variety	Multidimensional Robot Attitude Scale

Table 3. Quality in Use Model for Software Assessments [ISO 2011c]

Category	Metrics
Effectiveness	Effectiveness
Efficiency	Efficiency
Satisfaction	Usefulness
	Trust
	Pleasure
	Comfort
Freedom from Risk	Economic
	Health and safety
	Environmental
Context Coverage	Completeness
	Flexibility

4.2 Software Quality

While robot systems are physical constructs, the mechanics and HRI are driven by software. Many of the metrics assessing the effectiveness of HRI are therefore actually evaluating the performance of software. Standards for evaluating software quality were introduced in ISO 9126-1:2001 [ISO 2001], which has since been superseded by ISO/IEC 25010:2011 [ISO 2011c]. ISO/IEC 25010 presents two models for evaluating quality: quality in use and product quality. The *quality in use* model evaluates the overall impacts and outcomes resulting from the use of a given system on different stakeholders. Such metrics include effectiveness, efficiency, overall satisfaction with the system, freedom from risk, and the degree to which these metrics hold for specific contexts (Table 3).

The *product quality* model is used to characterize a given system based on its properties, including functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability (Table 4). Each of these characteristics areas define sub-characteristics that are assessed as having their requirements met satisfactorily or unsatisfactorily (or assessed on a Likert-like scale) as experienced by primary, secondary, and indirect users. Some topical areas such as functional suitability, reliability, and usability may be easily evaluated by the lay user, but many of the remaining characteristics and sub-characteristics may require subject experts to evaluate properly [Valenti et al. 2002]. Moreover, evaluations from different stakeholders are not easily combined or directly comparable due to the stakeholders' individual priorities and perspectives for assessing different system qualities. Such quality metrics are therefore subjective and non-repeatable.

4.3 Interfaces

There is a natural tendency to gravitate toward contextually nebulous terms such as "easy to use" or "intuitive" to describe HMI. While there exist a number of best practices for guiding the development of such interfaces—many of which will be discussed in this section—there is not a universal metric by which one interface may be directly compared with another. Moreover, because terms such as "easy to use" are ultimately subjective in nature, the practice of interface design is frequently addressed as being more of an art than a science, though it is heavily influenced by scientific studies such as psychology [Card et al. 1983] and physiology [Wickens et al. 2004]. As such, most metrics regarding the performance of interfaces are focused on the user. There is a balance that must be struck with the desire to have an interface that is easy or intuitive to use, but also provides a rich (i.e., full-featured) user experience.

Table 4. Product Quality Model for Software Assessments [ISO 2011c]

Category	Metrics
Functional Suitability	Completeness
	Correctness
	Appropriateness
Performance Efficiency	Time behavior
	Resource utilization
	Capacity
Compatibility	Coexistence
	Interoperability
Usability	Appropriateness recognition
	Learnability
	Operability
	User error protection
	UI aesthetics
	Accessibility
Reliability	Maturity
	Availability
	Fault tolerance
	Recoverability
Security	Confidentiality
	Integrity
	Non-repudiation
	Accountability
	Authenticity
Maintainability	Modularity
	Reusability
	Analyzability
	Modifiability
	Testability
Portability	Adapability
	Installability
	Replaceability

One of the earlier HMI evaluation methodologies was proposed by Roberts and Moran [1983] to evaluate the impact of the interface/tool—specifically, a text editor—itself. The methodology breaks the evaluation into four metrics: learning time, expert use time, error cost, and functionality (see Table 5). Learning time reflects the time necessary to first learn how to use the system, while expert use time captures the amount of time to perform some task assuming mastery of the tool. Error cost measures the time penalty when errors are made, and functionality is a measure of how full-featured the system is, and is calculated as

$$functionality = \frac{|F|}{|S|}, \tag{1}$$

where $F \subset S$ is the set of functions supported by the system of the larger set of all functions S. These metrics provide a fair (albeit incomplete) basis for comparing interfaces with similar goals and may be broadly applied to the collaborative robot use case. Capturing functionality,

Table 5. Interface Evaluation Metrics Proposed for Text Editors [Roberts and Moran 1983]

Metric	Measurement
Learning Time	Time (in seconds) for novices to learn how to perform basic editing tasks using the system
Expert Use Time	Required time (in seconds) for an expert to perform the basic editing tasks using the system
Error Cost	Time cost (in seconds) associated with making, discovering, and correcting for errors, and then resuming productive work
Functionality	Percentage of the complete taxonomy of tasks supported by the system

however, is not straightforward. In terms of industrial robot interfaces, the functional goals of robot systems are broadly scoped and are as varied as the number of unique designs, configurations, and applications of said robot systems. As such, attempts to define command taxonomies (e.g., Robot Operating System [Edwards and Lewis 2012] and the Canonical Robot Command Language [Proctor et al. 2016]) occasionally run the risk of gravitating toward the extremes of the spectrum ranging from *too cumbersome and specific*, to *too small and generic*. Defining a comprehensive, functional definition of robots, however, is beyond the scope of this report and is left as an exercise for a formal working group.

While such metrics are useful for capturing the overall effectiveness of the interface, they do not always address the specific aspects that contribute to the user experience. Many of the aspects that make a particular interface more intuitive for someone over a different interface could be considered a matter of personal taste. However, there are several best practices identified that, when ignored, tend to negatively impact the user experience. Many such design principles for user interfaces (UIs) are outlined in texts specifically designed for practitioners (e.g., Johnson [2014]; McKay [2013]). The quantifiable aspects of these principles are outlined in this article presently.

The two primary purposes of a UI are information presentation and system/process control. During normal manufacturing operations, information presentation will be the dominant function of the interface, but this information is provided (“presentation”) with the assumption that it will be leveraged for doing something useful (“response”). There are two metrics associated with these presentation-response expectations. The first is whether the response was correct (assuming that a correct response exists). The correctness value can be quantified in several ways, including (1) the number of erroneous responses, e; (2) the frequency of incorrect responses, $\frac{e}{\tau}$ over some time period, τ; (3) the mean time to an incorrect response, τ_{error}; (4) the measurable distance (if the response is numerically based), $\overline{\hat{r}r}$, from the given response, r, to the correct response, $\hat{r}$; or (5) some measurable consequence or cost, c, of an error (e.g., number or severity of manufacturing part errors, monetary value of waste, or system downtime). Comparing two different UI implementations as a function of the response correctness provides a quantifiable performance indicator of the interfaces.

The second metric is the amount of time lapsed between the initial display of information and the moment the operator confirms their response to the stimuli. Depending on the response desired, timing performances should be theoretically bounded to the timeframe outlined in Table 6. These timeframes provide lower limits, $\tilde{t}$, to the amount of time required for a given stimulus-response cycle. Large deviations from $\tilde{t}$ indicate either the interface design is not optimal for the application or the presentation of information is incorrect for the expected response.

There exist a number of timing indicators that may demonstrate the interface design may not be optimal for the given task. For instance, if the layout of controls is broadly distributed, operators may require additional time to complete certain tasks. Fitts’ Law [Fitts 1954]—which has been

Table 6. Important Timeframes for Human Responses

Range (s)	Perception and Cognitive Function	UI Timeframe
0.01–0.5	Perception of number of 1 to five items (0.05 s/item)	Quickly counting fewer than five items [Card et al. 1983]
	Perception of cause and effect	Allowable delay in hand-eye coordination [Card et al. 1983]
	Flinch reflex (0.015–0.4 s)	Minimum reactive motion response to stimuli [Davis 1984]
0.5–1.0	Attentional "blink" (0.5 s)	Required wait time after presenting important information before presenting more [Card et al. 1983]
	Reaction time for unexpected events (0.2–1.0 s)	Time after presentation of visual information that it is retained physically [Card et al. 1983]
6.0–30.0	Unbroken concentration on a task	Completing one step of a multi-step task [Card et al. 1983]
100.0+	Critical decisions in emergency situations	All information required for a decision is provided/found within this time [Johnson 2014]

successfully applied across a variety of different input mechanisms—states the time, t_{pos}, to move the hand to a target of size S that is a given distance, D, away is

$$t_{pos} = l_M \log_2 \left(\frac{D}{S} + 0.5 \right), \tag{2}$$

where $l_M = 100[70 \sim 120]$ msec/bit. Farther distances result in longer hand traversal time and more time searching for the appropriate control. As a general best practice, objects of interest should be placed prominently in the center of the field of view to minimize search time. Both data and control representations should be compact and visual (e.g., using iconography), with controls offering a select number of predefined potential responses rather than open-ended fields for text or variable manipulation. Similarly, controls and feedback mechanisms should be proximally placed near where current focus is. Examples of this practice include placing messages where operators are already looking, marking errors in red (text or boxes) or with symbols [Johnson 2014].

Another large influencer of performance timing are the gains earned through repeated use of an interface. The Power Law of Practice [Snoddy 1926] effectively outlines that, as one becomes more accustomed to the interface, the time t_n to perform a task on the nth trial follows a power law

$$t_n = t_1 n^{-\alpha}, \tag{3}$$

where $\alpha = 0.4\,(0.2 \sim 0.6)$. Specifically, repetitions of the same task should take less time given practice and competence. If no such gains are observed, it is possible the UI design or layout is interfering with the operator's performance.

While it may be convenient to associate poor interface performance only with the design choices, the presentation of information, if not appropriate for the application or expected response, may also negatively impact performance. The goal-oriented human thought cycle [Card et al. 1983] is fairly straightforward:

(1) Form a goal,
(2) Choose and execute actions necessary to make progress toward that goal,
(3) Assess the impacts of the selected actions, and
(4) Repeat steps 2 and 3 until the goal has been achieved.

People notice things more when they are specifically related to the specified goals and may be susceptible to change blindness [Simons and Levin 1997] or inattentional blindness [Simons and Chabris 1999] unless pertinent information is intentionally brought to the person's attention. Likewise, information may be ignored if it is deemed unrelated to the specified goals. The Gestalt theory of perception [Wagemans et al. 2012a, 2012b], for instance, gives a holistic perspective of the temporal and spatial relationships inferred by the presentation of visual objects. This Hebbian-like association of objects clustered or moving together enables people to draw conclusions in the absence of structure or information and fill in gaps in the context. As such, if information is not presented in a way that promotes this association (e.g., cluttered displays, spatially disparate indicators, or poor management of foreground/background messages), related information may become disassociated.

Similarly, information that is not presented in an intuitive or easily consumed manner may take longer to process. For example, while spoken language is naturally generated and understood verbally, writing and reading take more effort [Sousa 2006]. As such, best practices include using clear and consistent words or phrases, using simple typefaces with large fonts, and presenting information on clean backgrounds [Johnson 2014].

Even these practices may not be sufficient if the information itself is difficult to understand or respond to. If higher-level cognitive processes are required for a response, the Uncertainty Principle [Card et al. 1983] states that the decision time t increases with uncertainty about the judgment or decision to be made:

$$t_{dec} = F_C H, \tag{4}$$

where H is the information-theoretic entropy of the decision and $F_C = 150\ (0 \sim 157)$ msec/bit. For n equally probable alternatives [Hick 1952],

$$H = \log_2 (n + 1)\,. \tag{5}$$

For n alternatives with different probabilities of occurrence, p_j,

$$H = \sum_i p_i \log_2 \left(\frac{1}{p_i} + 1 \right). \tag{6}$$

While the time frames given in Table 6 can be considered targets under ideal circumstance, the operator's performance under real-world conditions is expected to be considerably longer. To this end, simpler, goal-oriented interface designs may improve task performance, but only at the expense of information completeness and process flexibility.

4.4 Communication Efficacy

A principal function of any user interface is the communication of relevant information to the machine operator. Similarly, the same interface must enable clear communication with the robot. This two-way communication is critical for effective HRI [Fong et al. 2006]. However, determining the efficacy of communications, or whether relevant information has been conveyed effectively, in real-time, is difficult. While communication efficacy is influenced by the quality of information being shared, it is ultimately reliant on the recipient to confirm that important data have been received and understood. The ability to establish common ground is complicated by the operator's assumption that a robot (particularly one that shares some human traits) will behave and respond as a human would [Jensen et al. 2017]. Methods and metrics for establishing common ground and situational awareness are described in more detail in Section 4.6. In this section, we focus instead on the metrics for determining whether information was communicated in a way that maximizes productivity with minimal loss of detail or relevance.

Historically, such information has been expressed either in text or via a sequence of lights. Yet, with advancements in processor and display technologies, information sharing is becoming increasingly visual. Moreover, with the robot's innate ability to move, information and meaning can also be conveyed through poses and gestures. Comparing the efficacy of information exchange through these various means can be a challenge.

When assessing the communication performance in the opposite direction (i.e., the human conveying information to the robot), one must be aware that the means by which information can be shared and expressed are nearly always significantly limited. Queries must be made in a way that the robot can understand and respond to; commands require explicit specificity to be executable; and requests for process status must be directly supported by programmed intelligence. All of this is system-, task-, and environment-specific, and anything that is not explicitly supported by the programmed sensing capabilities of the robot will be ignored. As such, confusion on the part of the robot system will be the result of syntax errors, not semantic or contextual ambiguity.

Ultimately, the only metric by which effective communication can be broadly and generically measured is communication time, t_{comm}. This is the amount of time required for a message to be generated, transmitted, and its meaning understood by the targeted recipient. This applies equally to text-based, verbal, and non-verbal communications, but capturing this time for non-text messages is nontrivial. As such, discussion in this section is exclusively focused on displayed messages for communication. Mechanisms for evaluating verbal and non-verbal (e.g., gestures, display lights, or sirens or warning tones) communications are not presented given the difficulty in capturing and characterizing interactions with operators.

Disruptions in communication adversely impact process (or work) performance, so the elements of communication delay, efficiency, and interruptions. Obviously, communications must be received on time if they are to be effective. Delays in transmission or comprehension reduce the remaining time to address issues before the loss of parts, equipment, or profit accrue. Traditionally considered a function of data quality, the measurement of timeliness is discussed in more detail in Section 4.5. Technological factors that may negatively impact communication effectiveness include transmission delays, jitter, and bandwidth [Steinfeld et al. 2006]. Such factors are relatively easy to isolate, characterize, and minimize.

An *efficient* message is one that is concise and to the point. Minimizing message length also minimizes the time necessary to transmit information, but may adversely effect communication time if the messages are not clear. As will be discussed in Section 4.5, there is a mathematically provable minimum amount of information necessary to convey information. This absolute minimum, however, is distinct from the minimum amount of information necessary to convey information clearly and without requiring additional processing time to translate (see Section 4.7 for the discussion on mental effort). An example of this distinction would be sending an error code (an absolute minimum amount of information) versus sending an error code accompanied by a brief description. As the operator becomes adept with the system, the descriptive text may eventually become superfluous.

Assuming the messages are expressed using the jargon or symbolism appropriate for the application and end-users, it can be assumed that the words themselves are already of an optimal length without sacrificing meaning [Piantadosi et al. 2011]. Thus, the content of the generated messages should be the principal focus of attention rather than the wording. When communicating critical information, we can borrow from the health-care field and formulate messages containing four principal components [Leonard et al. 2004]:

(1) a synopsis of the situation,
(2) the relevant background of the system or process,

(3) a diagnosis of the situation, and
(4) a recommendation of steps to move forward.

Messages using this structure provide all of the information for an operator to correctly respond to issues as they arise.

In providing all information without delay, it is possible to disrupt process cadences by sending too many low-priority messages when they are not critical [Rennecker and Godwin 2005]. Interruptions, or unscheduled events that cause work to be stopped, disrupt processes and negatively impact organization. They may also be effectively used to avert more costly interruptions (e.g., operator injuries or equipment failures) down the line. Confusing the high- and low-priority events may overload the cognitive abilities of operators to filter out the events that require immediate attention from those that can be safely addressed later. It is recommended the latter category be rescheduled or reformatted to minimize the impacts on productivity [O'Conaill and Frohlich 1995]. Factors that may influence the magnitude of impact—either positive or negative—include the number, frequency, length, and relevancy to the current task [Rennecker and Godwin 2005]. It is also suggested that different individuals' control motivations may result in different responses to external interruptions and may actually take steps to initiate or avoid process interruptions [Rennecker and Godwin 2005].

Tools for documenting, parsing, and evaluating such communication interactions have been developed (e.g., Burke et al. [2004]) that require expert analysis manual tagging to parse and label video logs. Such logs can provide the basis for post-factor analysis to assess back-and-forth flows of information, disruptions of process flow, situational awareness (see Section 4.6), mental effort (see Section 4.7), and task performance (see Section 4.10). As advances in natural language processing, machine learning, and machine vision improve, it is anticipated such tools may be automated.

4.5 Information Quality

Robotic systems are constantly giving feedback, mostly in the form of robot's physical state (e.g., cycle times, Cartesian pose of the tooling, and joint angles and temperatures). Most of this feedback is not consumable by an operator. In a thoroughly integrated workcell, additional operator-centric feedback may be conveyed to express the state of the manufacturing process. For feedback from a robot system to be useful to an operator, it must consist of timely, actionable intelligence. Specifically, feedback must be given when it is most relevant to a desired response, and it must consist of enough information that both the meaning and the expected response to the feedback are clearly understood. Good quality information from the robot to the operator enables the operator to make more informed decisions. Similarly, for situations where user input is required (e.g., Fong et al. [2003]; Kaipa et al. [2016]), high-quality feedback from the operator improves the robot's performance.

One of the largest contributing factors of the usefulness of reported information is the quality of said information. High-quality information contributes to the operator's situation awareness (see discussion in Section 4.6), whereas low-quality information detracts from it. "Low quality" information may be identified as insufficient or incorrect feedback during a collaborative task, but overloading the operator with superfluous information or providing unclear feedback can also be detrimental. The metrics for information quality—or "data quality"—vary somewhat according to application domain, but a generally agreed-upon model [Knight and Burn 2005] consists of 20 primary attributes, shown in Table 7.

In general, the principal factors that contribute to information quality are believability, value-added, relevancy, accuracy, interpretability, understandability, and accessibility [Wang and Strong 1996]. However, the weights of these attributes may shift based on the needs of the application

Table 7. Dimensions of Information Quality Reported by Knight and Burn [2005]

Category	Attribute	Description
Accuracy	Accuracy	Correctness of information
	Believability	Credibility of information
	Completeness	Information required for a given task is not missing
	Objectivity	Impartiality of information
	Reliability	Trustworthiness of the information
	Reputation	Quality of regard for information source
	Quantity	Volume of information available for use
Relevancy	Relevancy	Applicability of information to the task
	Timeliness	Presentation of information in a favorable amount of time
	Usability	The information's ease of use
	Usefulness	Contribution of the information to the task
	Value-added	Benefit gained by the information's use
Representation	Conciseness	Information is minimally represented yet complete
	Consistency	Repeatable format and compatibility with prior information
	Efficiency	Ease by which the task's information needs are met
	Understandability	Ease by which information is comprehended
Accessibility	Accessibility	Ease of information retrieval
	Availability	Speed of information retrieval
	Navigation	Ability to find and link to information
	Security	Access to information is correctly restricted

Italicized attributes indicate elements that are purely subjective in nature [Batini et al. 2009].

and operational environment. Almost all of these factors are operator-dependent and defined with respect to the task. This is consistent with the discussions in Section 4.3 and Section 4.6: The subjective preferences and internal processing of information by the operator contribute greatly to the performance of the task. As such, factors such as how the data are formatted [Miller 1996] may ultimately influence the operator's perspectives on data quality.

All of the attributes described in Table 7 may be assessed through subjective operator surveys. Given this subjective nature, quantifying data quality may be rather difficult. One approach to quantification is to simply assemble and assess feedback from multiple qualitative sources. Pipino et al. [2002], for instance, focus on metrics for objectively assessing data quality (as defined by the 20 attributes in Table 7) based on subjective perception assessments from all applicable stakeholders. Feedback from stakeholders are enumerated (e.g., per a Likert scale) and then evaluated by simple mathematical operations such as basic ratios of undesired results over the total results (i.e., $(outcome_{undesireable}/outcome_{total}) - 1$) for factors such as completeness or consistency, *min* and *max* operations, and weighted averages. However, it should be noted that many attributes have quantifiable elements that provide objective analogues for the subjective measurements. The survey by Batini et al. [2009] provides an excellent overview of the data quality literature leading up to 2009.

For information to be of any value, it must be accurate. Depending on the nature of the information, *accuracy* may be measured in a variety of ways. For instance, if the information not only needs to be correct but also delivered to the correct recipient, accuracy may be measured as the number of correctly delivered data packets [Jeusfeld et al. 1998]. Otherwise, if there is a quantifiable element to the data, accuracy may be measured as the L^2 norm of data errors, E,

$$accuracy = ||E||. \tag{7}$$

Alternatively, a more common metric is to measure the *syntactic accuracy* (e.g., see Batini and Scannapieco [2006], De Amicis and Batini [2004], English [1999], Falorsi et al. [2003], Loshin [2001], Pipino et al. [2002], Scannapieco et al. [2004], Su and Jin [2006], and Wang [1998]), which is effectively the ratio of the size of the set of correct responses, C, to the size of all responses received, N (i.e., the *quantity* of information):

$$accuracy = \frac{|C|}{|N|}. \tag{8}$$

The *believability* (or credibility) of the information being provided is often a second-guessing of the information accuracy. Without redundant sources of information or extensive testing to prove correctness of the information source, users may have a metered level of distrust with the data. This is particularly true when the user is unfamiliar with the equipment, process, or sources of information. Are the presented data accurate or subject to noise? Is it an absolute value or based on some baseline value? For example, some distance-measuring sensor systems express significant degradation of accuracy as a function of distance from the sensor. Without having intimate knowledge of the system's limitations, there is no reason to believe one measurement is more or less accurate than another.

To this end, believability is often considered a subjective measure in the absence of additional information. However, from the literature, one objective means of assessing believability is to assign default values to all incoming information and measure changes against this baseline. As such, the believability may be represented as either a raw count of the number of default values [Jeusfeld et al. 1998], $|D|$, or as a percentage of all non-default information:

$$believability = 1 - \frac{|D|}{|N|}. \tag{9}$$

This measure of believability may be considered a re-instantiation of the measure of *completeness*, which compares against a baseline consisting of null responses, S [Batini and Scannapieco 2006; De Amicis and Batini 2004; Lee et al. 2002; Loshin 2001; Pipino et al. 2002; Scannapieco et al. 2004; Su and Jin 2006; Wang 1998]:

$$completeness = \frac{|S|}{|N|}. \tag{10}$$

An alternative measure of completeness, however, evaluates the quantity of information actually received compared with the quantity of information expected, X [Jeusfeld et al. 1998; Lee et al. 2002]:

$$completeness = \frac{|N|}{|X|}. \tag{11}$$

Related to—but sometimes antithetical—the measure of completeness is the evaluation of information *conciseness*, which is assessed in terms of the length of the information sent compared with the concepts shared. For example, if two messages contain the same amount of information, but the first message conveys the information with 10% fewer bytes than the second message, then it is considered more concise. Measuring conciseness in this manner is not an absolute, however, so it has also been proposed to evaluate conciseness as the quantity of highly complex data structures, messages, or pages of information [Batini et al. 2009]. Similarly, *efficiency* is measured as a function of the amount of information required (e.g., number of data values, message length, or number of instructions), R, versus the total amount of information received [Pipino et al. 2002]:

$$efficiency = \min\left(\frac{N}{R}, \frac{R}{N}\right). \tag{12}$$

In the context of system and process diagnostics, functional performance information is expected to be conveyed in a timely, actionable manner. As such, factors such as *timeliness* play an important role in information quality. Timeliness, as measured by Ballou et al. [1998], is impacted by both how up-to-date (or *currency*) the information is as well as the expected useful shelf life (or *volatility*) of said information. The *currency* of information is measured simply as

$$currency = \left(t_{delivery} - t_{input}\right) + age, \tag{13}$$

where $t_{delivery}$ is the timestamp at which the information was delivered to the operator, t_{input} is the timestamp at which the information was requested (either manually or automatically), and *age* is the length of time separating the generation and transmission of the information. The timeliness of the information is thus measured as

$$timeliness = \left(\max\left[\frac{1 - currency}{volatility}, 0\right]\right)^s, \tag{14}$$

where s is a sensitivity factor of the volatility ranging from marginally sensitive, ($s < 1$), to highly sensitive, ($s > 1$). An alternative metric measures timeliness as a percentage of informative messages received within the required time frame, A [English 1999; Loshin 2001] (i.e., $t_A \leq volatility$):

$$timeliness = \frac{|A|}{|N|}. \tag{15}$$

Relevancy and *usefulness* are interconnected when evaluating information quality. The relevancy is a measure of the applicability of the information received and is Boolean in nature in that the information is considered relevant if any part of it is used. The usefulness of the information, however, is a measure of unused/non-applied information, U, or, conversely, the percentage of the information that is applied:

$$usefulness = 1 - \frac{|U|}{|N|}. \tag{16}$$

Another possible measure of evaluating relevancy and usefulness is to take the two as a combined value and estimate it as a measure of *response time* [Shah and Breazeal 2010]; specifically, evaluating the communication by the lapse in time between when the information was expressed and when a response was observable.

In the context of information quality, *consistency* refers to the adherence of guidelines for information formatting or representation. Such guidelines may specify data formats or ranges of expected values. The measure of consistency, therefore, would be a function of the number of values that are semantically valid, Q [Pipino et al. 2002; Wang 1998]:

$$consistency = \frac{|Q|}{|N|}. \tag{17}$$

An alternative approach would be to measure the diversion itself from the semantic guidelines (i.e., coding violations) as a distance from the expected value [Jeusfeld et al. 1998].

Finally, the quality of information access is a functionality of four measurable attributes: accessibility, availability, navigation, and security. *Accessibility*, or the ease of information retrieval, may be measured simply as the amount of time (or steps) required to identify and retrieve needed information. An alternative metric, similar to the measure of timeliness, evaluates accessibility in terms of time against a set deadline, $t_{deadline}$ [Eppler and Muenzenmayer 2002]:

$$accessibility = \max\left(0, 1 - \frac{t_{delivery} - t_{input}}{t_{deadline} - t_{input}}\right). \tag{18}$$

Separate from the retrieval time, *availability* is measurable as the time required for the system to respond to (i.e., initiate the retrieval process) the request for information [Eppler and Muenzenmayer 2002]. The measure of how difficult it is to discover what information to retrieve (or how to retrieve it), *navigation*, is characterized by the number of hierarchical steps [Eppler and Muenzenmayer 2002], the cost of missteps, and the number of circuitous routes possible to get to the same information. And *security* refers to the strength of the credentials required to access the information, or, conversely, the number of weak credentials present in the system [Eppler and Muenzenmayer 2002].

Assuming the quality of incoming data can be measured (i.e., by means of quantifiable errors), Ballou et al. [1998], using their information manufacturing model, give additional quantifiable measures for data quality and value to the customer. They define the data quality for output $y = f(x_1, \ldots, x_M)$, given M data unit inputs x, as some function of the measure of deficiencies in input variables (*component*), and the measure of errors introduced by processing the data ($effectiveness \in [0, 1]$):

$$quality_{out}(y) = f(component, effectiveness), \tag{19}$$

where *component* is the measure of the deficiencies of input variables

$$component = \frac{\sum_{i=1}^{M} w_i quality_{in}(x_i)}{\sum_{i=1}^{M} w_i}, \tag{20}$$

where

$$w_i = \left|\frac{\partial f}{\partial x_i}\right| |x_i|, \tag{21}$$

are the weights of the individual data components. Value to the customer (v) is generally defined as a function of the intrinsic value v_0, *timeliness*, and $quality_{out}$. An example of such a function is given as

$$v = v_0 \left(w\,(quality_{out})^a + (1 - w)\,t^b\right), \tag{22}$$

where $w \in [0, 1]$ is the customer's trade-off importance weight between quality and timeliness, a is a sensitivity to quality, and b is a sensitivity to timeliness. Here, v_0, w, a, and b are all user-defined.

4.6 Situation Awareness

The study of SA is broad and the available literature extensively vast. Many efforts to define and measure SA have focused on specific domains (e.g., military operations and response efforts) that are beyond the context of this article. Rather than providing a full survey of the topic here, we direct the reader to more comprehensive texts for evaluating SA by Endsley and Garland [2000], and for design strategies for SA by Endsley and Jones [2016]. Instead, this section is dedicated specifically to the measurement of SA within the context of manufacturing, and any analogues drawn from these other domains will be indicated as such.

For the sake of consistency with the established literature, we adopt the widely cited three-level SA model presented by Endsley in 1995 [Endsley 1995b], as shown in Figure 2. In this model, Level 1 SA represents the capacity to perceive and measure elements within the operational environment (e.g., people, work objects, and obstacles). Level 2 SA indicates the comprehension of the elements'

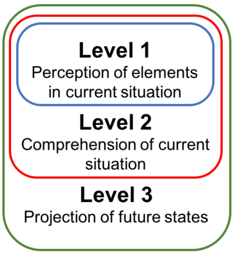
Level 1
Perception of elements in current situation
Level 2
Comprehension of current situation
Level 3
Projection of future states

Fig. 2. The three nested levels of SA as described in Endsley [1995b]. Each successive layer extrapolates greater awareness based on the previous layer's understanding of the world.

meaning within the operational context. And Level 3 SA demonstrates the ability to predict with some level of accuracy the elements' states (e.g., location and orientation) in the future. The stacked nature of this model states that Level 3 SA is impossible without also having Level 1 and Level 2 SA. If, at any point, Level 1 SA is lost, Level 2 and Level 3 SA are also forfeited.

HMI and HRI that fully support SA are marked as having fewer errors or requiring few corrective actions. Loss of SA is typically associated with missing critical information [Jones and Endsley 1996] and is attributed to poor system performance [Stanton et al. 2001]. While ultimately still being susceptible to black swan events [Taleb 2007], effective interfaces and interactions express greater safety and performance robustness against such unforeseeable situations.

In industrial settings, human-factors SA topics are traditionally focused on ensuring safety in the face of uncertainty (e.g., Cormier et al. [2015]). However, the support of the human operator's SA in manufacturing environments is equally important to maximize optimal process efficiency and minimize downtime. For instance, it has been observed that operators in autonomous manufacturing environments are more passive when they monitor the processes and equipment and are thus less responsive to system/process failures [Endsley and Kiris 1995]. To combat this, it has been proposed that reducing the autonomy of the system to actively engage the operators will summarily increase operator SA [Kaber and Endsley 2004]. This position is often in direct conflict with the desire to maximize autonomy to increase robot effectiveness and enable a single operator to support multiple machines (e.g., Crandall et al. [2005]). In such circumstances, however, the target is not SA in collaborative tasks, but is instead the maximization of equipment utility. The effectiveness of a robot's autonomy ("neglect tolerance"), J, is measured as a function of neglect and interaction time, the complexity of the environment (Q), and the interface (I) [Crandall et al. 2005]:

$$J\left(I,Q,t\right)=\begin{cases} J_N\left(I,Q,t_{off}\right) & \text{neglected} \\ J_S\left(I,Q,t_{on},t_N\right) & \text{otherwise} \end{cases}. \tag{23}$$

Here, J_N is the performance curve of the robot while the robot is operating autonomously, and J_S is the performance curve of the robot while it is being interfaced with by the operator. The timing profiles, t, consist of the time elapsed since the operator turned their attention toward the robot (t_{on}), the time elapsed since the operator turned their attention away from the robot (t_{off}), and the length of time the robot had previously been neglected (t_N).

A number of metrics have been proposed to measure the SA of an individual during an event. Born principally for military and aviation applications, assessment metrics such as the Situation Awareness Rating Technique (SART) [Taylor 1990], the Situational Awareness Rating Scale (SARS) [Waag and Houck 1994], and Situation Awareness Global Assessment Technique (SAGAT) [Endsley 1988, 1995a] are subjective in nature with experiences recorded as post-factor,

Table 8. The SA System Metrics from Salerno et al. [2005]

Dimension	Metric	Definition/Purpose
Confidence	Precision	Percentage of correct alerts
	Recall	Probability of detection
Purity	Misassignment	Percentage of evidence incorrectly associated
	Evidence recall	Percentage of detected events
Cost Utility	Cost utility	Percentage of cost savings
Timeliness	Time	Time between event and alert

pseudo-quantitative measures on unidimensional scales. Such metrics are generally well-regarded with advantages and disadvantages to their applications [Charlton 2002; Jones 2000].

Quantitative measures are less common given the highly personal and subjective nature of SA. As such, such metrics attempt to infer SA by measuring correlated, externally observable variables. For instance, the Situation-Present Assessment Method (SPAM) [Durso et al. 1998] attempts to quantify SA in real-time by measuring response delays as a function of the subject knowing where to look for information within the environment. Similarly, it is believed that, as a person's SA increases, the amount of additional information needed to complete a task decreases [Salerno et al. 2005]. A data-to-information ratio equates data consumption with the value of the operator's SA of objects or events:

$$\textit{Data-to-Information} = \frac{Number\ of\ Observations}{Number\ of\ Events}. \tag{24}$$

In addition to this data-driven evaluation, Salerno et al. [2005] define several performance-based metrics for evaluating the systems specifically designed to generate and maintain SA for emergency situations (see Table 8). Such metrics include the measurement of "*confidence*"—essentially, a measurement of false-positive and false-negative errors—which is the juxtaposition of "*precision*" (classification accuracy)

$$Precision = \frac{Number\ of\ Correct\ Detections}{Number\ of\ Detections} \tag{25}$$

and "*recall*" (detection accuracy)

$$Recall = \frac{Number\ of\ Correct\ Events\ Detected}{Number\ of\ Events\ in\ Ground\ Truth}. \tag{26}$$

For systems that trigger event notifications based on the accumulation of flagged data, the "purity" (or quality) of the data classification is measured as the scale of noise in data flagging (a "*misassignment rate*") and recall at the evidence level. The misassignment rate is calculated as

$$Misassignment\ Rate = \frac{\sum_{i=1}^{M} I\left(m_i\right)}{\sum_{i=1}^{M} S\left(m_i\right)}, \tag{27}$$

where M is the total number of events detected, $I\left(m_i\right)$ is the amount of data incorrectly tagged for event m_i, and $S\left(m_i\right)$ is the total amount of data associated with the event m_i. And *evidence recall* is evaluated as

$$Evidence\ Recall = \frac{\sum_{i=1}^{M} C\left(m_i\right)}{\sum_{i=1}^{M} E\left(m_i\right)}, \tag{28}$$

where $C(m_i)$ is the amount of data correctly tagged to belong to event m_i, and $E(m_i)$ is the collection of data in the ground truth for event m_i. There are also metrics to measure the time required by the system to bring an event to the operator's attention ("*timeliness*") and consequences of missing or incorrectly identifying events ("*cost utility*"). Here, cost utility is defined as

$$Cost\ Utility = \frac{\sum_{i=0}^{M} Cost\left(m_i\right)}{\sum_{j=0}^{N} Cost\left(n_j\right)}, \tag{29}$$

where the $Cost(\cdot)$ function is the cost (positive or negative) of a given event, $m_1 \ldots m_M$ are all events detected, and $n_1 \ldots n_N$ are the events that should generate operator notifications.

In terms of designing for SA, Endsley and Jones [2016] provide a comprehensive account of the threats to maintaining SA across multiple application domains and the principles to consider when creating UIs (primarily military and aviation applications, but also applicable to manufacturing tasks) to minimize the likelihood and impacts of these threats. The principal threats, they argue, are:

- focusing attention on specific features or aspects, while losing track of surrounding events or processes;
- relying on limited, volatile memory for activities that exceed the storage lifespan;
- stressors such as workload, anxiety, and fatigue, which limit the ability to process incoming information;
- information being presented at a faster rate than at which it can be processed;
- less-important information overriding more important information that needs the operator's attention;
- too many sources of information, which make it difficult to process everything and maintain SA;
- relying on incomplete or incorrect understandings of how to interpret incoming information;
- automated features that disengage operators, making it less likely to catch and correct errors.

Endsley and Jones further provide 50 design principles to combat these threats and support SA for specific requirements such as displaying process uncertainty, minimizing complexity, and raising alarms. These principles include organizing the presentation of data around operator goals rather than the technology, supporting operator comprehension by providing Level 2 SA information directly, providing information about sensor reliability, and keeping the interfaces goal-oriented by limiting the number of additional (potentially distracting) features. Factors supporting the design of automation systems are discussed further in Section 5.2.

In contrast to SA for human operators, Steinfeld et al. [2006] argue that SA for robots is reduced to two forms of awareness: human awareness and self awareness. Human awareness consists of the robot's ability to recognize, characterize, and adapt to humans. Human recognition refers to the functions relating to the identification, localization, and tracking of people. Human characterization, in contrast, attempts to capture human actions, intention, and attention. And human adaptation includes recognizing and adjusting to human behaviors and feedback.

The robot's self awareness is an internal assessment metric by which the robot monitors and evaluates its state. Such states include measures of the robot's limitations and capabilities (e.g., Shneier et al. [2015]), monitoring its own system health and process performance, and identifying and recovering from fault [Steinfeld et al. 2006]. Because of the complex interactions with machine tools and other robots, however, it is expected that there would be a level of external redundancy in system health monitoring, process performance, and fault detection.

Both the human and self awareness measures from Steinfeld et al. [2006] are qualitative in nature, but may be quantitatively assessed using external observers to verify the identification and modeling of humans and robots. In the case of human awareness, human-oriented perception is the combination of identification and localization of people in the workspace as measured against some ground truth [Shackleford et al. 2016]. Identification ~~, *id*,~~ is measured as an accuracy measure,

$$identification = \frac{|d_c|}{|d_{tot}|}, \tag{30}$$

where d_c is the number of correct "human" classifications/detections, and d_{tot} is the number of all detections made. Localization is the average accuracy of the robot determining where in three-dimensional Cartesian space the detected humans are:

$$localization = \frac{\sum_i^{d_c} ||g_i - p_i||}{d_c}, \tag{31}$$

where p_i and g_i are the measured and ground truth locations of the ith detected person, respectively.

In cases where no ground truth is available, however, a more probabilistic approach may be required to assess the system's *confidence* in its identification and localization accuracy. There are myriad probabilistic approaches within the literature for identifying and locating people, each of which have their own measures for evaluating the confidence of the assessment. Often, this is a measure of a quality of fit, typically bounded by the number of features, conditions, or constraints satisfied by the identification/localization algorithm. This can be reported as a simple ratio of *observed* versus *expected* features or constraints:

$$confidence = \frac{observed}{expected}, \tag{32}$$

or as a confidence interval in the form of $(\mu_{observed}, \sigma_{observed})$ if the observations themselves are probabilistic in nature. In such cases, the mean and standard deviation of observed conditions may either independently or collectively influence the reporting of measurement confidence (e.g., a given measurement may be assigned a low confidence if the number of observed conditions is low or the standard deviation is large). Measurement uncertainty (i.e., noise) distributions may also be leveraged to the same effect and have the added benefit of being useful for compensating for measurement noise using statistical estimators (e.g., Kalman filters).

The robot's quantifiable self awareness is a function of its ability to measure and model itself and consists largely of assessments of action performance and fault handling. Action performance, as a basic level of functionality, is essentially positioning and trajectory accuracy and repeatability. Additional software and sensors may be integrated for higher-level performance assessment and validation (e.g., see Section 4.10), but such capabilities are task-specific and not typical of a stock robot platform. Nominally, accuracy and repeatability are measured as a function of the relative Euclidean distance between the nominal and measured barycentric positions, $\hat{p}_i$ and p_i, respectively. Positional accuracy [ISO 1998] is measured as

$$A_P(i) = ||\hat{p}_i - p_i||, \tag{33}$$

and the orientation accuracy [ASTM 2014] may be measured as

$$A_O\left(i\right) = cos^{-1}\left(\frac{\text{trace}\left(\hat{R}_i R_i^{\text{T}}\right)}{2}\right), \tag{34}$$

where $\hat{R}_i$ and R_i are the 3×3 rotational matrices defining the nominal and measured poses, respectively. The repeatability may be measured as

$$R_P = \bar{A}_P + 3\sigma_P, \tag{35}$$

where

$$\bar{A}_P = \frac{\sum_i^n A_P\left(i\right)}{n} \tag{36}$$

and

$$\sigma_P = \sqrt{\frac{\sum_i^n \left(A_P\left(i\right) - \bar{A}_P\right)}{n-1}}. \tag{37}$$

Fault handling, unlike action performance, is a post-factor analysis of the robot's ability to detect, isolate, and recover from errors. Fault detection is an instantiation of syntactic accuracy given in Equation (8):

$$detection = \frac{|C_{found}|}{|N|}, \tag{38}$$

where C is the number of faults correctly identified, and N is the total number of faults identified. Similarly, fault isolation is a quantification of the number of errors that are correctly diagnosed and localized, $C_{diagnosed}$:

$$isolation = \frac{|C_{diagnosed}|}{|N|}. \tag{39}$$

And fault recovery is a measure of the number of conditions correctly moved from an error state to an error-free state, $C_{corrected}$:

$$correction = \frac{|C_{corrected}|}{|C_{diagnosed}|}. \tag{40}$$

4.7 Mental Effort

Within the manufacturing context, the goals of HRI are targeted at reducing mental and physical fatigue, increasing sense of control or contribution to the task, and ultimately reducing any interaction dynamics that might negatively impact production goals. The metrics for measuring such subjective impacts are largely qualitative in nature and tend to target impacts on the individual by assessing cognitive or stress loads. One widely applied qualitative metric [Hart 2006] is the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) [Hart and Staveland 1988], which performs a post-factor rating of the levels of cognitive and physical demands placed on the participants during a given task. These demands (shown in Table 9) are rated on different Likert scales and combined with "Source of Workload" weights and are intended to characterize both the task and the level of personal effort exerted to complete the task.

Alternative metrics include the Workload Profile (Table 10 [Tsang and Velazquez 1996]) and the Subjective Workload Assessment Technique (SWAT, Table 11 [Reid and Nygren 1988]). The Workload Profile evaluates multiple quality/demand elements of several mental dimensions on sliding scales in the range of 0 (no demand) to 1 (full effort), while the SWAT metric has participants select from among three possible responses for time, mental effort, and stress-load dimensions. Based on

Table 9. The Qualitative Metrics of the NASA-TLX [Hart and Staveland 1988] Ask Participants to Rank Mental Demands along Multiple Dimensions

Title	Scale	Description
Mental Demand	Low–High	Level of mental and perceptual activity required to complete the task.
Physical Demand	Low–High	Level of physical activity required to complete the task.
Temporal Demand	Low–High	Amount of pressure felt as a result of the pace of the task.
Performance	Good–Poor	Assessment (either internal or external) of the quality of work performed toward the task goals.
Effort	Low–High	Amount of work exerted (mentally or physically) to achieve the level of performance.
Frustration	Low–High	Level of internal stress (insecurity, discouragement, irritation, stress, or annoyance) felt during the task.

Table 10. The Qualitative Metrics of the Workload Profile [Tsang and Velazquez 1996] Ask Participants to Rate Their Experiences to the Different Dimensions on the Scale of 0 (No Demand) to 1 (Full Attention)

Dimension	Task Quality
Processing Stage	Perceptual/Central
	Response
Processing Code	Spatial
	Verbal
Input	Visual
	Auditory
Output	Manual
	Speech

a study by Rubio et al., the NASA-TLX is recommended for assessing impacts on individuals, while evaluations of cognitive demands are better evaluated using Workload Profile [Rubio et al. 2004].

As with any qualitative rubric, the subjective measures of the NASA-TLX, Workload Profile, and SWAT metrics may be influenced by any number of external stimuli that may vary as a function of time. Moreover, such metrics are not applicable when assessing the effectiveness of robotic tools or coworkers, since they do not feel fatigue or stress. Quantitative physiological measurements may be used to generate approximations of mental effort and fatigue in lieu of the more-common feedback models, but the mechanisms for taking such measurements are invasive and are not conducive with safety and performance requirements of a manufacturing environment.

As such, alternative quantitative metrics may be required to add context to (or be used in lieu of) the qualitative measures. Such metrics are effectively aimed to capture observable activity within a particular group and may be as simple as measuring the sharing of information (e.g., the number, frequency, and modality of messages exchanged [Guazzini et al. 2012]), capturing the lengths of time spent in the shared workspace or task, proximity of humans and robots throughout the task [Guazzini et al. 2012], or evaluations of collaborative team size and rates of change [Klug and Bagrow 2016; Palla et al. 2007].

Table 11. The Qualitative Metrics of the Subjective Workload Assessment Technique [Reid and Nygren 1988] Evaluates Tasks by Selecting from among Three (“Low,” “Medium,” “High”) Responses

Load Dimension	**Response**
Time	Often have spare time. Interruptions or overlap among activities occurs infrequently or not at all.
	Occasionally have spare time. Interruptions or overlap among activities occur infrequently.
	Almost never have spare time. Interruptions or overlap among activities are very frequent or occur all the time.
Mental Effort	Very little conscious mental effort or concentration required. Activity is almost automatic, require little or no attention.
	Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention is required.
	Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.
Stress	Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.
	Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.
	High to very intense stress due to confusion, frustration, or anxiety. High extreme determination and self-control required.

Mental effort may be approximated by observing the nature and experience (E) of a team of size M for a given project R worked on by user i

$$E = \frac{1}{M}\sum_i |R_i| - 1, \tag{41}$$

diversity

$$D = \frac{|\bigcup_i R_i|}{\sum_i |R_i|}, \tag{42}$$

and number of leaders (team members who contribute more than others in team k, where L_{ij} is 1 if user i is the lead of team j, zero otherwise)

$$L_k = \sum_{i=1}^{M_k} \min\left(\sum_j L_{ij}, 1\right), \tag{43}$$

of teams [Klug and Bagrow 2016] based on the number of projects/tasks a person works on, R_i.

4.8 Physical Effort

In Section 4.5, reliability and trustworthiness were introduced as a qualitative metric of information quality as assessed by the operator. A study by Sadrfaridpour et al. [2016] introduced a “trust model” metric of reliability based on the human’s performance, p_h, compared with the robot’s performance, p_r, at time t. The purpose of this model is to identify when the operator’s activity is disproportionate with the robot’s. The underlying theory is that a sign of trust between human and machine is when the two are equally active at the same time, rather than relying on turn-taking or an unequal distribution of work. This trust model is calculated as

$$trust\left(t\right) = c_1 trust\left(t-1\right) - c_2\left|p_h\left(t\right) - p_r\left(t\right)\right| - c_3\left|p_h\left(t-1\right) - p_r\left(t-1\right)\right| + r_{recover}, \tag{44}$$

where c_1, c_2, and c_3 are user-defined, positive constants, and $r_{recover}$ is a trust recovery rate. At time t the robot's performance is assessed based on its current level of activity normalized by its maximum possible activity. Here, $0 \leq p_r\left(t\right) \leq 1$, where 0 indicates the robot is idle, while 1 means the robot is moving at maximum permissible speed.

In contrast, the human's performance is based on their physiological capacity to work:

$$p_h\left(t\right) = \frac{f_{max}\left(t\right) - f_{thresh}}{f_{init} - f_{thresh}}, \tag{45}$$

where f_{max} is the maximum applicable muscle force, which diminishes over time, and f_{init} is the maximum force applicable at rest. The fatigue and recovery model of f_{max}, borrowed from Fayazi et al. [2013]:

$$f_{max}\left(t\right) = f_{max}\left(t-1\right) - c_f f_{max}\left(t-1\right)\frac{f\left(k-1\right)}{f_{init}} + c_r\left(f_{init} - f_{max}\left(t-1\right)\right), \tag{46}$$

where c_f and c_r are the fatigue and recovery constants unique to every person, respectively, and $f\left(t\right)$ is the applied force at time t. The value, f_{th}, is the threshold force applicable by a muscle and is equal to the smallest possible values of f_{max}, calculated as

$$f_{thresh} = f_{init}\frac{c_r}{2c_f}\left(-1 + \sqrt{1 + \frac{4c_f}{cr}}\right). \tag{47}$$

As a model of trust, this metric has a number of issues, not least of which being that it assumes both human and machine must be actively working to illustrate a level of trust. For example, turn-taking in a shared work environment, or a person holding down a workpiece while the robot performs some operation on it, or instances where the robot works separately from the human due to safety concerns (e.g., the robot moves a part to a separate work station to perform a quick welding operation) do not relate in any way to the human's trust of the machine. However, the measures of p_h and p_r do provide useful insights into the physical effort exerted by both humans and machines and are thus worth mentioning as a measure of teaming performance.

An alternative model of muscle exertion is presented by Peternel et al. [2016] as an index value at time t:

$$v_{index}\left(t\right) = 1 - e^{-\int \frac{g(t)}{c}dt}, \tag{48}$$

where $V\left(t\right) \in \left[0, 1\right]$ is the index value, $g\left(t\right) \in \left[0, 1\right]$ is an estimate of the current activity dynamics (e.g., an applied force or stiffness model or directly measured via Electromyography sensors), and C is an encoding of an individual's specific muscle parameters:

$$c = -\frac{g_{ref}t_{ref}}{\log\left(1 - 0.993\right)}. \tag{49}$$

Here, g_{ref} is a reference effort to measure the amount of time, t_{ref}, that a given muscle could endure.

These models of effort are occasionally used to assess ease-of-use for force-guided human-robot collaborative tasks such as material handling (e.g., Flixeder et al. [2016]; Fujii et al. [2016]; Villani et al. [2018]) and manipulation of components (e.g., Ficuciello et al. [2015]). In many instances, however, interactions between humans and robots are strictly non-contact for safety purposes. In such applications, verbal or gesture-based queues are used to communicate with the robots. For these, physical effort would effectively reduce to vocal strain [Zhao et al. 2018] and ergonomics

[Aromaa et al. 2018]. Alternative interfaces such as augmented or virtual reality may have additional physical effort factors associated, including neck strain from the weight of displays [Garrett et al. 2018] and cybersickness from extended use [Kennedy et al. 1993; Rebenitsch and Owen 2016].

4.9 The Human Response

The effectiveness of a robot's design, the performance of the interface, and the quality of information shared can all be for naught if the person working with the robot reacts negatively to the robot. A significant portion of the acceptance of the human-robot teaming may ultimately depend on personal preference, which is notoriously difficult to quantitatively capture *a priori*. Factors such as likability and the perception of intelligence cannot be accurately measured objectively at all. As such, it is necessary and unavoidable to include subjective measures that capture the operators' feelings toward the robotic equipment with which they are working.

The challenge, however, is to determine the minimum subset of reaction and impression attributes that succinctly capture the human experience. Recall from Section 4.1 the discussion of multiple surveys/questionnaires in Table 2. In that table, 41 different subjective measures are presented, each of which are captured in different ways by different surveys. Some of these attributes are similarly captured in other subjective metrics discussed earlier, yet many are specific in targeting the human response.

One of the more popular subjective measures for measuring operator reactions to robots is from the Godspeed Questionnaire Series (GQS) developed by Bartneck et al. [2009]. The GQS captures human responses across a sliding scale for five categories: anthropomorphism (how human-like is a robot's appearance or behavior), animacy (how lifelike is a robot's appearance or behavior), likeability (positivity of first impressions), intelligence (acting or reacting intelligently), and safety (awareness of hazards to the human). These categories, shown in Table 12, are based on an extensive survey of sociology and psychology metrics and capture a user's personal assessment of a robot's capacity to interact with people. While ultimately focused exclusively on assessing how a robot looks or acts, the GQS is a useful tool for determining whether the person responds positively or negatively to design choices for the evolution of the robot and its interface(s).

In contrast with the GQS's evaluation of appearance, realism, and intelligence, metrics of trust take the assessment of robot competence one step further to determine whether the person working with the robot has faith in the robot's capabilities. In Section 4.8, a proposed objective measure was presented in Equation (44) that equated simultaneous activity levels with inferences of trust. However, this is just one aspect of trust and does not necessarily correlate with an individual's comfort level or faith in the abilities of the robot to perform as expected. For example, the Multi-Dimensional Measure of Trust [Ullman and Malle 2018] asks users to rate robots on their perceived capabilities, ethics, sincerity, and reliability as factors of trustworthiness. The Almere model [Heerink et al. 2010], in contrast, specifies trust as a factor of personal integrity and reliability. Joosse et al. [2013] refer to trust as a measure of credibility as measured by the Source Credibility Scale [McCroskey et al. 1973] from 1973, which captures credibility as a function of competence, character, sociability, extroversion, and composure. In contrast, the Negative Attitudes Toward Robots [Nomura et al. 2006] scale measures a person's trust in robots in general, reducing interactions to hypothetical situations and encounters, social influences of robots, and the attribution of emotions (Table 13). Another metric of trust measurement was proposed by Muir that establishes trust as a combination of predictability, dependability, faith, competence, responsibility, and reliability [Muir 1994]. For a more comprehensive discussion on trust metrics and the impacts that different factors have on trust, we refer the reader to the survey by Cho et al. [2015].

Also worth noting here is the System Usability Scale developed by Brooke [1996]. Within the HRI research domain, this is frequently used in lieu of the product quality model from [ISO 2011c]

Table 12. Human Response Categories from the Godspeed Questionnaire Series [Bartneck et al. 2009]

Category	Subcategory
Anthropomorphism	fake vs. natural appearance
	machine-like vs. human-like behavior
	unconscious vs. conscious behavior
	artificial vs. lifelike appearance
	moving rigidly vs. moving naturally
Animacy	dead vs. alive appearance
	stagnant vs. lively actions
	mechanical vs. organic movement
	artificial vs. lifelike appearance
	inert vs. interactive behavior
	apathetic vs. responsive behavior
Likeability	user likes vs. dislikes appearance
	perceived friendly vs. unfriendly
	perceived kind vs. unkind
	perceived pleasant vs. unpleasant
	perceived awful vs. nice
Intelligence	incompetent vs. competent
	ignorant vs. knowledgeable
	irresponsible vs. responsible
	unintelligent vs. intelligent
	foolish vs. sensible
Safety	anxious vs. relaxed
	agitated vs. calm
	inactive vs. surprised

for measuring usability. The System Usability Scale is a 10-item survey that assesses operator preferences along the lines of intent to use, function integration, ease of learning, and system consistency. As with the NASA-TLX, the System Usability Scale is not specific to robotics. Moreover, it finds broad use and applicability in both industrial and non-industrial applications.

4.10 Task Performance

When integrating any new robot system for a specific application, one must take into account the impact of the system and its interface on the task performance. From a subjectively qualitative perspective, task performance can be inferred based on survey responses to inquiries regarding team cohesion (i.e., how everyone felt the team performed as a collaborative unit), distribution of workload, quality of work, and so on. Such evaluations may be useful for assessing whether the task (and the team members) benefit from the collaborative teaming.

Given that the purpose of integrating automation into a task is to increase product quality and process effectiveness, efficiency, and throughput, the quantitative metrics for evaluating the impacts on task performance are associated with manufacturing performance indicators. From a resource perspective, the effectiveness and efficiency of the team (and, subsequently, the individual contributors and robotic equipment) are indelibly tied to these performance indicators. *Effectiveness*, or the quantification of how well the task is completed, may measure any number of different

Table 13. The Negative Attitudes toward Robots Scale from Nomura et al. [2006] Assesses Attitudes along Three Hypothetical Categories

Subscale	Item
S1: Negative attitude toward situations of interaction with robots	I would feel uneasy if I was given a job where I had to use robots.
	The word "robot" means nothing to me.
	I would feel nervous operating a robot in front of other people.
	I would hate the idea that robots or artificial intelligences were making judgments about things.
	I would feel very nervous just standing in front of a robot.
	I would feel paranoid talking with a robot.
S2: Negative Attitude toward Social Influence of Robots	I would feel uneasy if robots really had emotions.
	Something bad might happen if robots developed into living beings.
	I feel that if I depend on robots too much, something bad might happen.
	I am concerned that robots would be a bad influence on children.
	I feel that in the future society will be dominated by robots.
S2: Negative Attitude toward Emotions in Interaction with Robots	I would feel relaxed talking with robots.
	If robots had emotions, I would be able to make friends with them.
	I feel comforted being with robots that have emotions.

manufacturing factors, but ultimately reduces down to two basic elements: (1) the number of completed operations (assembly, packaging, inspection, etc.), o_t, and (2) the number of damaged, lost, or scrapped parts (i.e., waste), o_w. The two can be combined into a single equation as

$$effectiveness = \frac{a\,(o_t - o_w) - (1-a)\,o_w}{o_t}, \tag{50}$$

where $0 \le a \le 1$ is a user-defined weight specifying the company preference of quantity versus quality. In contrast, *efficiency* measures the performance as a function of resource utilization for a given operation. Common metrics for this include the *average time* required to complete a task over a given period of time, t_{comp},

$$t_{avg} = \frac{t_{comp}}{o_t}, \tag{51}$$

or its inverse, *throughput*:

$$throughput = \frac{o_t}{t_{comp}}. \tag{52}$$

Optimizing for the time element consists of either minimizing t_{avg} or maximizing *throughput*. Either of these are sufficient for capturing efficiency in many application domains (e.g., response robotics [Steinfeld et al. 2006]). However, in manufacturing, we must also take into account the minimization of waste as a factor for efficiency. As such, we define our measure of efficiency as

$$efficiency = \left(1 - \frac{o_w}{o_t}\right)\left(\frac{o_t}{t_{comp}}\right) = \frac{o_t - o_w}{t_{comp}}. \tag{53}$$

Fluency, as described in Hoffman [2019], provides an additional metric for timing efficacy and synchronization between human-robot team members. Fluency includes both subjective survey questions and objective timing measurements. The objective measures consist of human idle time H_{IDLE}, robot idle time R_{IDLE}, concurrent activity time C_{ACT}, and functional delay F_{DEL}.

For a specific collaborative task instance $\langle H, R \rangle$ over some arbitrary time period, the n periods of human activity are denoted as

$$H \equiv \{h_i\}_{i=1}^{n} \equiv \{(s_{h_i}, d_{h_i})\}_{i=1}^{n}. \tag{54}$$

Here, period i is denoted as h_i, which starts at time s_{h_i} and has duration d_{h_i}. The robot's sequence of activity over m periods is represented as

$$R \equiv \{r_j\}_{j=1}^{m} \equiv \{(s_{r_j}, d_{r_j})\}_{j=1}^{m}, \tag{55}$$

where s_{r_j} marks the start of the jth period sequence, and d_{r_j} is its duration. The total task time is represented as

$$T = \max\left(s_{h_n} + d_{h_n}, s_{r_m} + d_{r_m}\right), \tag{56}$$

where the human always starts the first activity at time $t = 0$.

Using the above descriptions, the percentage of the total task time where the human was not active, H_{IDLE}, can be described as follows:

$$H_{IDLE} = 1 - \frac{1}{T} \sum_{i=1}^{n} d_{h_i}. \tag{57}$$

Similarly, R_{IDLE}, the percentage of total task time where the robot is not perceivably active, can be described as

$$R_{IDLE} = 1 - \frac{1}{T} \sum_{i=1}^{m} d_{r_i}. \tag{58}$$

It should be noted that the "perceivable" nature of R_{IDLE} can be interpreted multiple ways depending on the environment and the nature of the interaction.

Concurrent Activity, or C_{ACT}, the percentage of total task time during which both agents are active concurrently, can be calculated as

$$C_{ACT} = \frac{1}{T} \left[\max\left(0, s_{h_1} + d_{h_1} - s_{r1}\right) + \sum_{i=1}^{n} \left(\max\left(0, s_{ri-1} + d_{ri-1} - s_{hi}\right) + \max\left(0, s_{hi} + d_{hi} - s_{ri}\right)\right) \right]. \tag{59}$$

Functional Delay, F_{DEL}, is described as the accumulated time between the completion of one agent's action and the beginning of the other agent's action and is represented as a ratio of total task time. The mathematical representation of F_{DEL} makes a simplifying assumption that actions do not accumulate more than once and is described as

$$F_{DEL} = \frac{1}{T} \sum_{i=1}^{n} \left(s_{r_i} - s_{h_i} - d_{h_i}\right). \tag{60}$$

As detailed in Hoffman [2019], the rates of these measures can vary when dealing with small values of m and n, and it may be more appropriate at times to evaluate these metrics as an average agent turn. In general, the user of these metrics should be aware of the variance in lengths of the time periods before choosing to provide the metrics as an average or as a series.

From a business perspective, arguably the single-most important metric for evaluating the effectiveness of any system is the return on investment (ROI). Traditionally, the measure of ROI has been a tool for evaluating material purchasing decisions as an assessment of the timeline to recuperate investment costs for purchase and integration (*cost*) as a function of expected financial gains (*gain*) over a given period of time (e.g., one year). ROI is calculated as

$$ROI = \frac{gain_{term} - cost}{cost}, \tag{61}$$

where $gain_{term}$ is the gross profit, and *cost* is the combination of equipment cost, integration fees, and expected/amortized expenses (e.g., maintenance and training). For some, however, the perspective of ROI has shifted slightly to be purely a function of profit margins over the foreseeable life of the equipment, $gain_{total}$:

$$ROI = gain_{total} - cost. \tag{62}$$

Neither metric is inherently superior to the other, and selection of one or the other is ultimately at the discretion of the end-user. It is also worth mentioning that certain costs may be assessed differently depending on the institution. For example, some companies may only consider charges associated with actively producing goods in the computation of ROI, whereas support costs (e.g., programming or maintenance) may be accrued as overhead.

If implementing an entirely new task, one must isolate the HRI/HMI from the the task, itself. This is made challenging, however, given the impacts of HRI/HMI on key task performance indicators such as mean time to failure,

$$\bar{t}_{fail} = \frac{\sum_{i=0}^{f} t_{fail,i}}{f+1}, \tag{63}$$

(where f is the number of failure events, and $t_{fail,i}$ is the time to failure for event i), setup time, t_{setup}, and cycle time, t_{cycle}. For a thorough measurement, variability must be introduced into the process by experimenting with different interfaces and interactions. This is not conducive to manufacturing environments, however, nor are the results of the experimentation generalizable. One may, however, estimate the effects by reducing the complexity of the task to known variables and inversely measuring the influence that adding the complexity had on the task performance. For example, in a multi-robot configuration [Marvel et al. 2018], a reduction in complexity would reduce a multi-robot setup to a single-robot setup. A subset of the metrics for the evaluation of the single- versus multi-robot setup are given in Table 14. As an example, consider the impact of process timing. How long would it take a single robot to complete a given task (including the time necessary for tool changes, reconfigurations, and the use of fixtures)? How long would a multi-robot configuration require to complete the same task? The difference is the net gain.

For the human-robot case, one would assess the task from the human-only perspective versus the robot-only perspective. Human-centric perspectives are typical for both assistive (e.g., Tapus and Mataric [2008]) and response (e.g., Burke et al. [2004]) robotic applications and technologies. In the manufacturing domain, the introduction of robots into a historically manual process (or, conversely, increasing the role of human operators in automated processes) would be directly impacted by the HRI and HMI. The cost of integrating human-robot collaborative teams is thus evaluated as an ROI calculation.

In terms of evaluating HRI/HMI effectiveness, however, the impacts on ROI are almost impossible to accurately quantify given the predictive nature of the equation. How, then, does one measure the impact of HRI/HMI on the performance of a task? The measure is necessarily reflective: One compares the performance of the system leading up to the change and then measures the performance after the change (e.g., see the discussion of the metrics from Roberts and Moran [1983] in Section 4.3). As such, a more effective metric would be to evaluate the impact of HRI and HMI on the compounded overall equipment effectiveness (OEE) [Godfrey 2002]:

$$OEE = availability \times performance \times quality. \tag{64}$$

Table 14. Partial Collection of Metrics for Measuring Multi-robot Assembly Performance from Marvel et al. [2018]

Metric	Category	Measurement
Efficiency	Complexity	Complexity of task strategy
		Impact of using alternative task strategies
		Single- vs. multi-robot task complexity
	Effort	Programming time duration
		Optimization/tuning duration
		Commissioning vs. use timing
		Single- vs. multi-robot program timing
		Single- vs. multi-robot cost
	Quality	Ratio of task successes to failures
		Mean time to failure
Timing	Process	Average time to complete a single task
		Maximum time to complete a single task
		Minimum time to complete a single task
		Standard deviation of task times
		Impact of using alternative task strategies
		Sensitivity to parameter changes
		Time spent performing task vs. other motions
		Time required for single- vs. multi-robot configurations to complete tasks

Here, *availability* is the percentage of scheduled time the equipment is available to operate

$$availability = \frac{scheduled\ time - lost\ time}{scheduled\ time}, \tag{65}$$

performance is the percentage of operational speed of the equipment

$$performance = \frac{actual\ machine\ speed}{design\ machine\ speed}, \tag{66}$$

and *quality* is the ratio of good to bad parts produced

$$quality = \frac{number\ of\ good\ parts}{total\ parts\ produced}. \tag{67}$$

The HRI/HMI may impact any of these three elements in different ways, so the impact on OEE is highly variable as a function of the interactions between human and machine. For instance, improvements made to the timely display of quality diagnostics information may reduce equipment downtime, thus increasing total equipment *availability*. Similarly, in supportive human-robot collaborations, variations in the operator's confidence in the machine's capabilities and responsiveness may impact total production throughput, impacting *performance*.

If the HRI/HMI is the only element that has changed, OEE is an effective tool to capture the impact of the collaboration. Otherwise, these metrics are merely indirect indicators of HRI/HMI effectiveness toward task performance, and changes may be correlative rather than causal for improvements in OEE. Drawing parallels to the actual collaboration between user and machine, an analog OEE metric may be calculated to determine the extent and effectiveness of the interactions for task performance.

Thus, for HRI/HMI efficacy, we can define the OEE of the interface as

$$OEE_{HMI} = \begin{array}{c} availability_{HMI} \times performance_{HMI} \\ \times quality_{HMI}. \end{array} \tag{68}$$

Here, we define $availability_{HMI}$ as

$$availability_{HMI} = \frac{scheduled\ time - lost\ time}{scheduled\ time}, \tag{69}$$

where *scheduled time* is the number of hours allocated to perform the task, and *lost time* is the amount of time spent instructing or correcting the machine's actions or behavior. We define $performance_{HMI}$ as

$$performance_{HMI} = \frac{collaboration\ speed}{solo\ speed}, \tag{70}$$

where *collaboration speed* is the amount of time required to collaboratively perform a task through the HRI/HMI, versus *solo speed*, which is the amount of time required to perform the task alone (either the operator or the machine). And we define $quality_{HMI}$ as

$$quality_{HMI} = \frac{interactions - corrections}{interactions}, \tag{71}$$

where *interactions* are the number of times the operator interacts (physically or audibly) with the machine during the period of observation, and *corrections* are the number of times the operator had to correct the machine's behaviors as a response to the interactions.

5 MANUFACTURING HRI/HMI DESIGN FRAMEWORK

In the previous sections, we introduced and discussed the nature and metrics of HRI and HMI. In this section, we seek to compose a general guidance framework for influencing the design and evaluation of collaborative, manufacturing robot HMIs to maximize their effectiveness for HRI.

While the framework discussed in this section may be used for iteratively improving interfaces and interactions to enable more effective collaborations, this is not the guiding intention. Rather, this framework is intended to reestablish how the research and development communities approach interface designs. In a collaborative team, the HRI and HMI must be leveraged to establish trust. The operator must trust that the robot will not cause harm, that it will behave predictably, that it will follow the guidance of the operator (short of causing harm to personnel, equipment, or processes), and that its responses to uncertainty will be safe and reasonable.

As with any evaluation methodology, the recommendations and subset of suggested metrics that follow should be considered a "superset" and may not be applicable for all HRI and HMI designs, applications, and integrations. A subset of test methods and metrics will thus constitute the actual evaluation of a given technology implementation. However, by providing a canonical collection of metrics, we aim to provide a means by which different technologies can be consistently and repeatably assessed. In this section, guidelines will be provided to aid in the selection and interpretation of relevant metrics.

5.1 Interactions between Metrics

It has been seen previously that different metrics may be interpreted in different ways. For instance, the equation for credibility (Equation (9)) can also be interpreted as a measure for completeness (Equation (10)). Similarly, we have seen multiple metrics and test methods for effectively measuring the same thing (e.g., timeliness in Equations (14) and (15)). To this end, when one is developing a test methodology for evaluating the effectiveness of HRI and validation of HMI, one should exercise care that the metrics are accurate, useful, and unique.

A number of common hazards associated with metrics and test methods include:

(1) Metrics that can be used and interpreted a multitude of different ways do not bring new information or may be subject to misinterpretation/misidentification.
(2) Metrics used in one domain may be meaningless or inappropriate for other applications (e.g., the number of verbal exchanges as a means of evaluating communication efficacy may not capture the full extent of interactions in noisy environments).
(3) Some measurements may negatively influence or interfere with other metrics (e.g., constantly asking participants to rate their experiences may distract from the evaluations of teaming performance in human-robot tasks).
(4) Subjective metrics/questions may be worded in a way that could lead subjects' behaviors or influence responses to subsequent measures.
(5) Questions that are phrased positively or negatively may also influence responses. As such, questions are best phrased neutrally or have the same questions asked numerous ways, including both positively and negatively to negate the influence of framing.
(6) The use of certain measurement tools may impact how people behave (e.g., cameras and wearables, which could make people uncomfortable or self-conscious of performance). The mechanisms by which sensing is shared (or hidden) may also impact the subjects' behaviors or performance (e.g., live displays of camera feed may make some people feel self-conscious).
(7) Some task metrics and data collection mechanisms may fatigue subjects, which could negatively impact results. For example, questionnaires that survey study participants too frequently may elicit responses reflective of the participants' reactions to the test, not to the robot or teaming conditions. Similarly, attempts to gain real-time feedback of qualitatively subjective measures may be more of a nuisance to the participants than an asset.

It is also worth noting that seemingly related metrics are not necessarily directly correlated. For example, a reduction in mental effort does not necessarily imply a more effective interface or higher task efficiency. Recall the discussion in Section 4.3 regarding interface metrics. Assume 15 different manufacturing tasks can be reduced to a single button click on a control panel based on simple classifications of process conditions. Given these 15 tasks are all that are expected to be supported, one can deduce from Equation (1) that the functionality is 1.0. Similarly, learning and expert use times are minimal. However, what if each button is located on a separate tab of a virtual control panel? The cognitive load is minimal, but the majority of the operator's time may be spent interacting with the inefficient interface, simply scrolling through tabs to find the appropriate button(s) to press. Similarly, the task could be so mundane that maintaining the operator's attention could be a significant challenge.

To this end, not all metrics are universally applicable or relevant. Evaluating HRI technologies is therefore highly dependent on unique properties of the tasks, interactions, and interfaces that shape the nature of the teaming between humans and machines. These aspects are summarized in Table 15 and discussed presently. As will be repeated in the next section, it is recommended that, when selecting test methods and metrics, major stakeholders should be consulted and their preferences documented to aid in the prioritization of evaluation criteria. These preferences can then also be used to judge when their respective needs are met and the designed interfaces and interactions are ready for deployment. One possible mechanism for capturing these stakeholder preferences is described in Weiss and Schmidt [2010].

When selecting metrics for evaluating HRI technologies, one must first establish *what the goals of the interactions/interfaces are to be*. Quantitative measures should reflect the performance goals of the task (such as OEE, ROI, and process key performance indicators), while qualitative measures

Table 15. Factors of Consideration When Selecting Metrics and Test Methods

Factor	Quantitative Properties	Qualitative Properties
Task Goals	Task performance metrics	Expected influencing factors for the user
Interaction Method	Quantifiable factors of usability	Factors that reflect operator preferences
Environmental Factors	Impacts of environment on human/robot/team performance	Impacts of environment on human/team comfort and well-being

should assess the potential impacts on the operator (including ease of use and quality of work). For example, if an interface is intended to facilitate system maintenance, quantitative measures for accuracy and response times are likely to be highly relevant, as are qualitative measures for usability and information/data quality. In contrast, if the interaction is intended to provide instruction or material support, quantitative measures such as learning time and efficiency and qualitative measures such as effort and operator awareness may be of higher priority.

Second, metrics should be commensurate with *the means by which people will be interacting with the robots.* This includes a consideration of how the interfaces are expected to be most frequently used. Quantifiable metrics include those that measure the aspects of the interfaces and interactions for which the manufacturers and integrators are responsible. These factors include safety, utility of mixed initiatives, and the number of features available. Qualitative measures should reflect those aspects of frequent use that impact the operator's comfort when working with the robots. Such factors may include overall workload and effort, reliability, trust in the system, anxiety, and the ability to understand presented information.

Third, metrics should be relevant to the operational and environmental conditions in which the systems will be used. For example, measuring audio quality is superfluous in noisy environments, while situation awareness is highly relevant to areas that are frequented by human traffic or hazardous work items. Quantitative metrics should be selected to accurately measure how the environmental and operational conditions can impact the performance of the human-robot teaming. These include safety, operator/robot awareness, and task/process time. In contrast, qualitative measures should be selected to evaluate the impacts of the environment on the operator's well-being. This can include measures for situation awareness, mental and physical effort, and ease of use.

5.2 HRI Design Recommendations

Although direct data does not yet exist for human-robot teams in manufacturing, a survey of the literature elicits some clues as to how effective HRI in manufacturing should work. Several recurring themes emerged in analyses of teleoperated human-robot teams that are equally applicable in co-located human-robot collaborations in industrial settings.

Just as the nature and intended application of interactions and interfaces impacts the selection of metrics and test methods, they also impact the overall design of the systems to be used. Whether effective or not, HRI designs are frequently custom-built for specific applications in specific environments. This is true for assistive robot technologies used in the home as well as collaborative robots used in manufacturing. However, such tasks and environments are highly variable between and within a facility.

While reviewing the test methods and metrics frequently used within HRI/HMI research, many truisms emerged that should be considered when designing new human-robot team systems. Many of these recommendations are specific to design considerations that, while not necessarily directly tied to all of the performance metrics described previously, do impact the user's experience and the quality of information provided.

(1) Whenever possible, include the major stakeholders throughout the design process of the HMI, as prescribed in ISO 9241-210:2010 [ISO 2010]. Operator roles, as discussed in Section 2.2, and considerations for safety (Section 3.3) should influence the selection of stakeholders. For general-purpose interfaces, such user-oriented interactions may not be possible. However, for purpose-built interfaces, ISO 9241-210:2010 provides valuable instructions for designing human-centered interfaces.

(2) Per the discussion in Section 4.3 regarding goal-oriented though cycles and change blindness, use automation of functions sparingly and target operations that require routine actions [Endsley and Jones 2016]. Such limitations require greater engagement of the operator and reduce operator response time in the event of critical situations (Table 6).

(3) Graphically provide the operator with timely, actionable information about the status of the robot and the task to establish situational awareness quickly and efficiently [Murphy and Burke 2005; Murphy et al. 2004; Riddle et al. 2005; Scholtz et al. 2004]. Graphical representations of information allow for faster digestion of useful information (Table 6). In general, visual presentation of information often results in the most timely reactions with fewer operator errors, while text slows the reactions of operators and may lead to more errors [Perrin et al. 2008]. Similarly, information should be presented in "natural" or "familiar" ways, as uncommon/novel or uncomfortable interfaces may prove less effective [Perrin et al. 2008]. Ultimately, it is vitally important to provide transparency in the automated tasks such that it is clear what the robots are doing and what they are trying to achieve [Endsley and Jones 2016]. Refer to Section 4.6 for more detail on assessing neglect tolerance and awareness.

(4) Enable human in-the-loop interference to ensure safety, mitigate unanticipated problems, and directly manipulate the world [Endsley and Jones 2016; Goodrich and Olsen 2003; Murphy and Burke 2005]. Refer to Section 2.2 for a discussion of the different potential access and interface requirements given the different roles of operators. Similarly, the HRI should reduce the amount of effort necessary for directly controlling the robot [Goodrich and Olsen 2003] (see Section 4.7 and Section 4.8 for discussions of mental and physical effort). This includes reducing the amount of attention and time required to address issues with the robot as opposed to focusing on the collaborative task. Easy or intuitive interfaces for robot control in manufacturing include hand-guiding techniques or canned recovery procedures.

(5) Reduce the level of additional training to use the robot whenever changes are made to the interface or task [Casper and Murphy 2003; Murphy and Burke 2005] (see Section 4.3 for a discussion on evaluating learning time). Given that operators tend to be cognitively overloaded (see Section 4.4) and are unlikely to be willing/able to take on additional responsibilities, accommodating changes made to the system should require little or no additional learning.

(6) Related to the previous point, prioritize or facilitate communications, tasks, and negotiations between collaborators to manage operator attention (to prevent cognitive overload) and to ease the burden on human operators to "spread the word" to other humans or robots [Gold 2009; Goodrich and Olsen 2003; Johnson 2014; Murphy and Burke 2005; Murphy et al. 2004; Riddle et al. 2005; Scholtz et al. 2004].

(7) Enable a single operator to effectively interface with or control multiple robots [Goodrich and Olsen 2003]. Specifically accommodating this functionality typically enforces system design criteria that benefit operator awareness and process flow. See Section 4.6 for information regarding enabling and assessing situation awareness in such circumstances.

(8) Enable the fusion of multiple (perhaps unreliable) sensors to provide information in an integrated form [Scholtz et al. 2004]. While merely providing more information is not necessarily guaranteed to improve task performance or situation awareness, it has the potential to benefit the operator in instances of uncertainty. See Section 4.3 for a discussion of the uncertainty principle and Section 4.6 for an overview of the impacts of uncertainty on situation awareness.

(9) Similarly, allow the operators to manipulate the way they receive information or control the robot (e.g., granularity and fidelity of information flow and level of control) [Goodrich and Olsen 2003]. This will enable the operator to have more direct control over how information is presented to support situation awareness, reduce learning time, and filter out unnecessary, potentially distracting data flows.

(10) Enable intuitive manipulation of the robot-world relationship (e.g., representation of poses/configurations, and manipulate registrations between coordinate systems) [Gold 2009; Goodrich and Olsen 2003; Scholtz et al. 2004]. Again focused on maintaining situation awareness, requiring the operator to mentally transform their world view to match that of the robot can increase cognitive load and may contribute to operational errors that could impact process performance or operator safety.

With these tenets in mind, it is evident that a collaborative robot may have multiple interfaces for interacting with human operators using different contextual modalities and providing information in different ways. There are no global prescriptions for what these modalities should entail, as their requirements and limitations will be subject to application limitations, environmental constraints, and user preferences. Natural language processing, for instance, is not effective in noisy environments, and hand gestures are not effective when equipment or work items requiring both hands to lift or control are used. Similarly, different people have different preferences or needs for button layout, accommodations for disabilities or protective equipment (e.g., gloves, visors/goggles, and ear protection), formatting data, color schemes, and so on. As such, the capacity for adaptability and customization may be key for on-the-fly optimizing an HMI for specific users or job functions (e.g., Micire et al. [2011]).

Given the functionality of diagnostics and software control, physical interfaces are dictated by the application domain. The display is the principal component and should be featured prominently. Keyboards may be needed for some inputs, but numerical keypads may be a preferable complexity reduction if the only keyed inputs are numbers. Ultimately, however, programming robots and correcting for errors is often easiest through direct manipulation of the end-effector. Interfaces such as force control or robot-mounted six-axis joysticks allow the user to more intuitively move the end-effector into place without concern about contextual (orientation or translation) changes based on where the operator happens to be standing.

5.3 Performance Metrics

To be a meaningful tool, a set of metrics must provide sufficient coverage of the key performance indicators (KPIs) of the systems under test. Though industry has a number of KPIs—many of which are focused on the plant level [Lindberg et al. 2015]—the model of metrics coverage in this article is based on individual, team, and task performance. Within the context of human-robot teaming, demonstrating completeness is made challenging given the variability and scope of the collaborative tasks in factory environments. In Section 5.3, several metrics were introduced that could be used to assess and assure the performance of HRI in industrial HRI. In this section, we attempt to provide a minimum viable subset of these metrics to help guide and evaluate the development of HRI and HMI technologies.

To enable effective collaboration [Marvel 2014b], robots must be able to:

(1) temporally and spatially coordinate their motions,
(2) comprehend the task to which they have been assigned,
(3) communicate high-level, task-relevant information,
(4) have situation awareness about themselves, their environment, their task, and the humans with whom they are working, and
(5) maintain safety in the face of uncertainty.

These properties establish the framework within which the metrics and test methods are expected to function. Each aspect, however, has a multitude of criteria that must be captured. For example, the paper published by Ma et al. [2018] specifies that the primary components of teaming include (1) communication, (2) coordination, and (3) collaboration. Communication is expected to convey awareness, intent, and state. Coordination involves not just temporal and spatial synchronization, but also the maintenance of common understanding, altering one's own actions to direct the task progression, and predicting the actions of the other members of the team. Collaboration, according to Ma, is specifically focused on the maintenance and support on shared knowledge, intention, and goals.

HRI performance is assessed at both an individual (operator and robot) and team level [Steinfeld et al. 2006] and frequently involves measures of both physical and mental interaction to estimate the impacts of active communications (e.g., response time, proximity, duration of contact, and information quality). From a group dynamics perspective, other important factors of interaction include elements of trust [Boies et al. 2015; De Jong et al. 2016; MacArthur et al. 2017], which is influenced by proximity and contact [Armstrong 2008; Hall 1963] and speed of approach [Butler and Agah 2001]. It is not coincidental that these measures are also the factors of safety as discussed in Section 3.3. Additional metrics of trust include the elements of utility of mixed initiative in teams [Freedy et al. 2007].

Another key factor of HRI performance is the assessment of how effective human-robot teams are at completing their tasks. Task-based KPIs of HRI performance typically include economic factors such as ROI and product quality, as justification is often necessary for choosing one manufacturing process over another. Time-based factors include OEE, throughput, mean time to failure, setup time, and cycle time [Kangru et al. 2018; Weiss et al. 2013; Zhu et al. 2018].

Given these KPIs, a sufficiently complete HRI metric must include the following:

(1) team performance,
(2) individual performance (operator and robot), and
(3) task performance.

The team performance must include metrics of team composition, cohesiveness (effectively, how well the human-robot team performs), and trust. Individual performance must include measures of situation awareness (for the self, partner, process, and environment), task contribution, effort, and safety. And task performance must include both economic and process performance indicators.

From an HMI perspective, the KPIs are effectively focused on the robot and the task. For the robot, the performance metric is completeness of features of the interface. The robot's interface will be task-specific and thus is attributed to the task's HRI KPIs. Similarly, process diagnostics and feedback are also task-specific and effectively capture the quality of information being exchanged between the human and the robot.

5.3.1 Objectively Quantifiable Metrics. A number of quantitatively objective measures have been presented in the preceding sections that, although disparate in their respective foci, provide a comprehensive perspective of HRI and HMI system effectiveness when combined. From these available metrics, we selected what we consider to be a minimal subset that best captures the quantifiable aspects most directly tied to HRI performance and present these in Table 16. These metrics reflect the impacts that HRI and HMI have on the performances of the team, individual contributors, and the process as a whole. In many cases, we encountered multiple metrics that effectively measure the same elements (possibly with a different label), but we also note that the data collection for some metrics also directly benefits the measures for other metrics. For instance, the measures for information timeliness (Section 4.5, Information Quality) are also largely reflected in the measure for communication time (Section 4.4, Communication Efficacy). Of all the possible quantitative measures available, the subset of measures given in Table 16 are largely measurable automatically, with many measurable in real-time. Moreover, because it is assumed that the collaborative dynamic was chosen deliberately, metrics that are biased toward pure manual or automatic operations (e.g., neglect tolerance) are not included.

Interface design may impact a collaborative effort in four possible ways. Foremost, the performance of the manufacturing process may be impacted by the initial decision to incorporate human-robot collaborative teams. ROI and OEE are dependent on not only the decisions and designs leading up to the initial implementation of the interfaces and interactions, but also the extent to which feedback and control are supported through the HRI and HMI. With limited or uncertain information flow, the optimum performance for a given process may actually limit the performance of the collaborative team. That said, given the nature of human-robot collaborative interactions, poor HRI and HMI obviously impact the performance of the team as a whole. From there, interfaces and interactions may influence the performance of the human operators and robot systems individually.

5.3.2 Subjectively Qualifiable Metrics. While much emphasis has been placed on the quantifiable measurements, the qualifiable human perspective should not be ignored. Although not exhaustive, highlights of these metrics, shown in Table 17, are intended to encapsulate the perspectives of major stakeholders of human-robot collaborative teams and provide bases for systematic and procedural improvements regarding the quality of the interfaces and interactions. Virtually all of the qualitative measures are accumulated through surveys to which respondents may provide feedback along a Likert scale. Thus, data may be compiled, tabulated, and reported in a quasi-quantitative manner. Such results may be weighted accordingly based on stakeholder priority and anonymized to account for confidentiality of feedback.

The volume of subjective measures and survey tools available is large, and the metrics employed by most research in HRI are rarely validated, and are customized for the specific application the researchers intend. It may be difficult to select the measures that most accurately capture the nature of the intended interactions between humans and machines. To this end, we refer the reader back to Table 1 for selecting which survey tools meet their particular needs.

That said, many of the metrics identified in these surveys are irrelevant to most manufacturing applications and environments. While these metrics may be important when selecting robots for domestic applications, factors such as animacy, pleasure, and sociability should be ignored for the industrial use case. Moreover, relatively few survey tools have been thoroughly evaluated for validity. As such, we cannot recommend that most survey tools be included in a standardized suite of measurement tools.

Table 16. Set of Quantitative Metrics for Objectively Measuring the Performance of HRI and HMI

Subject	Category	Metrics [references] (Equations)	Section
Team	Performance	Effectiveness (Equation (50))	4.10
		Efficiency (Equation (53))	4.10
	Utility of mixed initiative	% Robot assistance requests [Steinfeld et al. 2006]	4.2
		% Human assistance requests [Steinfeld et al. 2006]	4.2
		Number non-critical human interruptions [Steinfeld et al. 2006]	4.2
		Communication time	4.4
Operator	Situation awareness	Reaction time (measured against timeframes in Table 6)	4.6
		Precision [Salerno et al. 2005] (Equation (25))	4.6
		Recall [Salerno et al. 2005] (Equation (26))	4.6
Robot	Self awareness	Self-monitoring and modeling (e.g., Equations (33) and (35))	4.6
		Fault detection, isolation, and recovery (Equations (38), (39), and (40))	4.6
	Human awareness	Human-oriented perception (Equations (30) and (31))	4.6
	Features	Functionality [Roberts and Moran 1983]	4.3
	Safety	Separation distance [ISO 2016]	3.3
		Approach speed [ISO 2016]	3.3
		Pressure [ISO 2016]	3.3
		Force [ISO 2016]	3.3
Process	ROI	Gain/profit	4.10
		Equipment/software cost	4.10
	OEE	Equipment availability (Equation (65))	4.10
		Performance (Equation (66))	4.10
		Quality (Equation (67))	4.10
	Interface	Learning time [Roberts and Moran 1983]	4.3
		Expert use time [Roberts and Moran 1983]	4.3
		Error cost [Roberts and Moran 1983]	4.3
	Timing	Setup time [Kangru et al. 2018]	4.10
		Cycle time [Kangru et al. 2018]	4.10
		Mean time to failure (63)	4.10
	Diagnostics and feedback	Information accuracy [Wang 1998] (Equations (7) and (8))	4.5
		Information accessibility (Equation (18))	4.5
		Timeliness (Equation (14))	4.5
		Usefulness (Equation (16))	4.5
		Consistency (Equation (17))	4.5

5.4 Example Case Studies

To illustrate the application of the metrics presented in Section 5.3, three case studies with varying degrees of human-robot interaction are presented from examples in the literature. Two case studies are borrowed from the 2012 survey of HRI in the automotive industry by Shi et al. [2012] and a third from the collaborative robot safety ontology paper by Marvel et al. [2015a]. Each case study

Table 17. Set of Qualitative (or Quasi-quantitative) Metrics for Subjectively Measuring the Performance of HRI and HMI

Subject	Category	Metrics [references]	Section
Team	Team composition	Team cohesion	4.10
		Workload distribution	4.10
		Diversity	4.7
		Leadership distribution	4.7
	Team performance	Quality of work [Steinfeld et al. 2006]	4.10
Operator	Workload	NASA-TLX [Hart and Staveland 1988]	4.7
		Workload Profile [Tsang and Velazquez 1996]	4.7
		System Usability Scale [Brooke 1996]	4.9
	Situation awareness	SAGAT [Endsley 1995a]	4.6
	Operator performance	Quality of work [Steinfeld et al. 2006]	4.10
Robot	Human awareness	Human modeling and monitoring [Steinfeld et al. 2006]	4.6
		Human sensitivity [Steinfeld et al. 2006]	4.6
	Robot performance	Quality of work [Steinfeld et al. 2006]	4.10
Process	Interface	Software quality [ISO 2011c], Table 4	4.2
		Quality in use [ISO 2011c], Table 3	4.2
	Diagnostics and feedback	Accuracy	4.5
		Availability	4.5
		Reliability	4.5
		Usability	4.5
		Understandability	4.5
		Value added	4.5

is described in detail, and the relevant metrics specified in Section 5.3 are identified based on the nature and degree of interaction.

5.4.1 Case Study: Low-interaction Welding Workcell. The state of practice within industry has historically demonstrated little, if any, actual interaction between humans and robots. This was primarily due to safety concerns. Workcells were monoliths of automation activity, where human operations were limited to specific stations on the periphery of the cell.

In the first case study from Shi et al. [2012], such a configuration is considered for HRI evaluation (Figure 3). In this example, a single human operator works with three industrial robots to complete a welding operation. Here, a human operator enters a safeguarded loading station, loads auto body parts into fixtures on a rotary table, and exits the station to initiate the robots' spot-welding motions. The robots do not move until the operator has both (1) exited the loading station and (2) pushed a button to start the robots' programs. Once the program begins, the table rotates into the welding configuration, and the two welding robots spot-weld the parts together. As the welding operation is being executed, the operator can load an additional fixture with parts for the next welding cycle. When the welding cycle concludes, a third robot removes the welded sub-assembly for transfer.

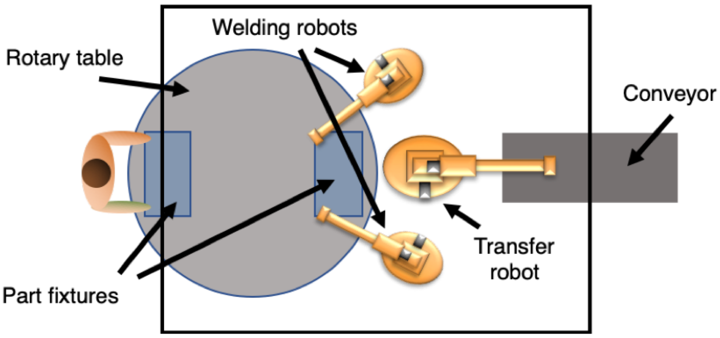

Fig. 3. In the low-interaction study, the operator interacts indirectly with the robots through a fixtured rotary table [Shi et al. 2012].

In this example, there is no direct interaction between the human and the robots. The rotary table acts as an intermediate interface, accommodating the part transfer from the human to the robots. As Shi et al. note:

> The hardware plays a role to prevent the direct interaction between the human and the robot by (1) enabling the human work outside the robot working envelope and (2) enabling the human to perform their own tasks asynchronously with the robot. However, the hardware transfer device, such as the rotary table or the input conveyor, adds cost, takes additional space, and makes the workcell less flexible for new product changes. [Shi et al. 2012]

Here, the collaboration is sequential, but the robots' actions are completely separate from the operator's tasks. Once the parts are loaded and the operator leaves the station, the human's role of the task is completed. This is very much a solitary application, so the team metrics (task performance and utility of mixed initiative) do not apply. Similarly, due to the absence of any direct interaction to impact the performance of the human or the robot, either physical or cognitive, the individual performance metrics are similarly irrelevant. This leaves only the process performance metrics as relevant here for the evaluation of HRI and HMI.

Depending on the operator's role in the control of the robots, the interface and diagnostics/feedback metrics may or may not be noteworthy. If the operator is responsible for starting, configuring, and/or monitoring the robot's programming or mechanical components, the interface and diagnostics/feedback metrics are critical for measuring the effectiveness of these limited interactions outside of the welding operations. However, if the operator who works with the robot during normal operations is not the same person responsible for maintaining the robots, these metrics are pertinent only for the maintenance operation. In either case, the ROI, OEE, and process timing metrics are important as KPIs for the workcell.

5.4.2 Case Study: Medium-interaction Grinding Workcell. With the 2012 revision ISO 10218 [ISO 2011a, 2011b] and the 2016 publication of ISO/TS 15066 [ISO 2016], direct interactions between humans and robots became viable solutions for manufacturing tasks. These interactions, however, see limited utility due to lingering impacts of the safety culture. As such, many tasks involving direct HRI in manufacturing are infrequent and of short duration.

In Marvel et al. [2015a], a hypothetical workcell is presented that involves the hand-off of assembled parts from the human operator to a robot for surface finishing at a stationary grinding station (Figure 4). In this example, the human operator is responsible for the assembly of parts,

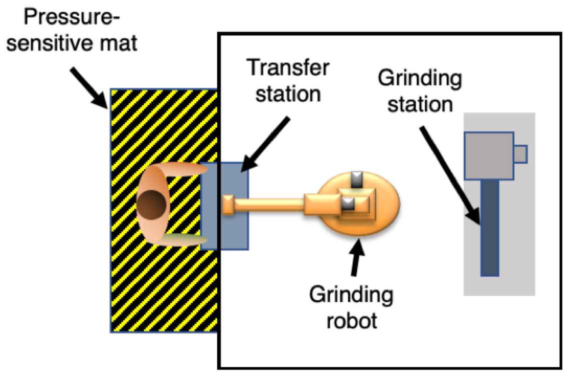

Fig. 4. In the medium-interaction case study, the operator interacts directly with the robot by means of a parts hand-off, but the grinding task is undertaken by the robot alone [Marvel et al. 2015a].

while the robot performs the grinding actions. This workcell combines the dexterity of a human operator with the repeatability of a robot. As a time- and cost-saving factor, the transition between the assembly and grinding processes is enabled by a direct hand-off from the human to the robot.

Away from the workcell, the human performs an assembly operation, completing a new subassembly. Once the assembly process is finished, the human takes the subassembly to the transfer station—the only location through which the human may interact with the robot—and activates the robot's program by stepping onto a pressure-sensitive mat. The robot moves to a predefined location within reach of the operator and waits for the operator to place the assembled part into its gripper. Once the part is in place, the operator signals the robot to close its gripper, which then actuates and corrects for minor positioning and orientation errors. The robot does not move again until the operator leaves the transfer station, stepping off of the pressure-sensitive mat. The robot then moves the part to the grinding station for surface finishing.

In this example, the only direct interaction between the human and the robot occurs during the hand-off event. The remaining operations separate the actions of the robot and the human such that they do not influence the other's performance. As such, the focus of this particular case study is on the hand-off activity.

During the hand-off, there is no direct communication between the human and the robot save that of the presence detection via the pressure sensitive mat. As such, for the team's metrics, utility of mixed initiative and team composition are not pertinent, leaving only the team's performance metrics (effectiveness, efficiency, and quality of work).

The hand-off process is driven by the operator and the robot's actions—save the time to move into position for the hand-off—have little impact on the operator's performance. As such, while the operator's metrics of situation awareness and performance are immaterial, the workload metrics are relevant for assessing the task's effect on the operator.

However, the operator's actions can directly impact the robot's performance. For instance, loading errors or delays may impact the robot's ability to perform the subsequent grinding task, so the robot's performance is critical to monitor. The ability to detect such issues necessitates the evaluation of self and human awareness. Because the robot's actions are not expected to change during the course of the operation, the evaluation of features is not pertinent. Similarly, because there is no direct physical interaction planned, the evaluation of pressure and force for safety are not required, though separation distance and approach speed are both important for maintaining a safe working environment.

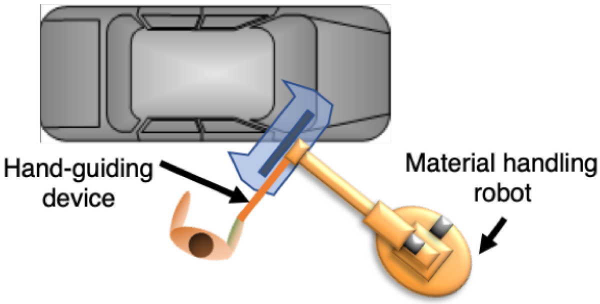

Fig. 5. In the high-interaction study, the operator interacts directly with the robot via a hand-guiding interface [Shi et al. 2012].

Moreover, given the limited interaction between the robot and the human, the assessment of the hand-off process is similar to that of the welding operation discussed previously. Specifically, the interface and diagnostics/feedback metrics are necessary only if the operator is responsible for programming and maintenance. Otherwise, only the ROI, OEE, and process timing are the important KPIs.

5.4.3 Case Study: High-interaction Assembly Workcell. While the use of direct human-robot interaction is still relatively minor, the frequency and duration of interactions are trending upward. This is due largely to the steady introduction and refinement of new collaborative robot safety functions. One of the more common applications of high interaction is collaborative assembly of large parts using hand-guiding. In Shi et al. [2012], one such use case is presented as an intelligent lift-assist robot for large, awkward, and/or heavy components (Figure 5).

In this example, the human operator directly controls the motions of the robot by means of a physical interface (e.g., a force/torque transducer or a six-degree of freedom joystick). This interface has the performance equivalence of a jogging function on a teach pendant, but the robot is running in automatic mode rather than a teach mode. This interface also enables the operator to control the robot's tooling. During these operations, the human must intentionally maintain robot activation (e.g., by means of a three-position enabling device), and the robot's speed is limited for safety purposes. With the exception of a "return to start" function activated upon the release of the enabling device, the robot is under the direct control of the operator at all times. During this return action, the human operator may work simultaneously near the robot on other tasks.

Because this operation requires the direct collaboration, physical contact, and synchronization of the operator and the robot, the robot and the human impact each other's performance and therefore the team's performance as a whole. Even when the robot is not under the operator's direct control, both are co-located and can influence one another. As such, all metrics for the team, operator, and robot are important for evaluation. Even if the operator is not responsible for the programming or maintenance of the robot, the nature of interaction requires a usable interface and clear diagnostics and feedback for both process efficiency and the maintenance of a safe working environment. Therefore, all process metrics are similarly relevant.

6 CONCLUSIONS AND FUTURE WORK

In this report, we presented an overview of HRI and HMI design considerations and the means by which their effectiveness can be assessed. The purpose of this effort is to advance innovations by providing a series of metrics by which HMI and HRI may be equally evaluated for their respective functions. By creating such a rubric, researchers and developers may optimize their designs for the user experience, enabling faster industry adoption, more intuitive interaction, and clearer process and system feedback. And by leveraging both quantitative and qualitative metrics, developers may target HMI and HRI improvements for both the broader user base as well as specific end-users. From the user's perspective, these metrics also provide a means by which they may select interfaces, configure interactions, or provide meaningful feedback to their technology providers such that their needs may be met.

Ongoing work at the U.S. National Institute of Standards and Technology is focused on the validation of the metrics and test methods for assessing human-robot teaming performance and to provide a quantitative measure for "intuition" of interfaces and interactions. These efforts include surveys of application domains in which human-robot teaming is particularly effective, HRI benchmarking and repeatability studies, and the execution of human trials for the control and situation awareness of robot actions using a spectrum of interface technologies. The goals of these continued efforts are to establish baseline test methodologies by which industry and academic partners may assess and drive forward advancements in HRI/HMI design.

DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this article to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

Alayne C. Armstrong. 2008. The fragility of group flow: The experiences of two small groups in a middle school mathematics classroom. *J. Math. Behav* 27, 2 (2008), 101–115.

Susanna Aromaa, Nikos Frangakis, Domenico Tedone, Juhani Viitaniemi, and Iina Aaltonen. 2018. Digital human models in human factors and ergonomics evaluation of gesture interfaces. *Proc. ACM Human-Comput. Interact.*, Vol. 2. ACM, 6.

ASTM. 2014. *ASTM E2919-14 - Standard Test Method for Evaluating the Performance of Systems that Measure Static, Six Degrees of Freedom (6DOF), Pose.* Standard. ASTM International.

L. Atzori, A. Iera, and G. Morabito. 2010. The Internet of Things: A survey. *Comput. Netw.* 54 (2010), 2787–2805.

Donald Ballou, Richard Wang, Harold Pazer, and Giri Kumar Tayi. 1998. Modeling information manufacturing systems to determine information product quality. *Manag. Sci.* 44, 4 (1998), 462–484.

Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Social Rob.* 1, 1 (2009), 71–81.

Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* 41, 3 (2009), 16.

C. Batini and M. Scannapieco. 2006. *Data Quality Concepts, Methodologies and Techniques.* Springer.

Kathleen Boies, John Fiset, and Harjinder Gill. 2015. Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity. *Leader. Quart.* 26, 6 (2015), 1080–1094.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psych.* 25, 1 (1994), 49–59.

John Brooke. 1996. SUS–A "quick and dirty" usability scale. *Usab. Eval. Ind.* 189, 194 (1996), 4–7.

P. Brooks. 2001. Ethernet/IP - Industrial protocol. In *Proceedings of the IEEE International Conference on Emerging Technologies in Factory Automation.* IEEE, 505–514.

Jennifer L. Burke, Robin R. Murphy, Dawn R. Riddle, and Thomas Fincannon. 2004. *Task Performance Metrics in Human-robot Interaction: Taking a Systems Approach.* Technical Report. DTIC Document.

John Travis Butler and Arvin Agah. 2001. Psychological effects of behavior patterns of a mobile personal robot. *Auton. Rob.* 10, 2 (2001), 185–202.

Fabrizio Caccavale and Ian D. Walker. 1997. Observer-based fault detection for robot manipulators. In *Proceedings of the IEEE International Conference on Robotics Automation*, Vol. 4. IEEE, 2881–2887.

Stuart K. Card, Allen Newell, and Thomas P. Moran. 1983. *The Psychology of Human-computer Interaction.* L. Erlbaum Associates Inc.

Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The robotic social attributes scale (rosas): Development and validation. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interactions*. ACM, 254–262.

J. Casper and R. R. Murphy. 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Trans. Syst., Man, and Cybern. - Part B: Cybern.* 33 (2003), 367–385.

Samuel G. Charlton. 2002. Measurement of cognitive states in test and evaluation. *Handbook Human Factors Testing Eval.* (2002), 115–122.

Jin-Hee Cho, Kevin Chan, and Sibel Adali. 2015. A survey on trust modeling. *ACM Comput. Surv. (CSUR)* 48, 2 (2015), 28.

Olivier St-Martin Cormier, Andrew Phan, and Frank P. Ferrie. 2015. Situational awareness for manufacturing applications. *Proceedings of the Comp. Rob. Vision* (2015), 320–327.

Jacob W. Crandall, Michael A. Goodrich, Dan R. Olsen, and Curtis W. Nielsen. 2005. Validating human-robot interaction schemes in multitasking environments. *IEEE Trans. Syst., Man, Cybernetics-Part A: Syst. Humans* 35, 4 (2005), 438–449.

Michael Davis. 1984. The mammalian startle. *Neural Mechanisms Startle Behav.* (1984), 287.

Fabricio De Amicis and Carlo Batini. 2004. A methodology for data quality assessment on financial data. *Studies Comm. Sci.* 4, 2 (2004), 115–137.

Bart A. De Jong, Kurt T. Dirks, and Nicole Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *J. Applied Psych.* 101, 8 (2016), 1134.

B. Doyle. 2015. North American Robotics Market Sets New Records in 2015. " http://www.robotics.org/content-detail.cfm/Industrial-Robotics-News/North-American-Robotics-Market-Sets-New-Records-in-2015/content_id/5951".

F. Driewer, M. Sauer, and K. Schilling. 2007. Discussion of challenges for user interfaces in human-robot teams. In *Proceedings of the Euro. Conference on Mobile Robotics*1–6.

J. L. Drury, J. Scholtz, and H. A. Yanco. 2003. Awareness in human-robot interactions. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*. IEEE, 912–918.

Francis T. Durso, Carla A. Hackworth, Todd R. Truitt, Jerry Crutchfield, Danko Nikolic, and Carol A. Manning. 1998. Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly* 6, 1 (1998), 1–20.

S. Edwards and C. Lewis. 2012. Applying the robot operating system (ROS) to industrial applications. In *Proceedings of the IEEE International Conference on Robotics Automation: ECHORD Workshop*. IEEE.

Mica R. Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE Aerospace and Electronics Conference* IEEE, 789–795.

Mica R. Endsley. 1995a. Measurement of situational awareness in dynamic systems. *Hum. Fact.* 37, 1 (1995), 65–84.

Mica R. Endsley. 1995b. Towards a theory of situational awareness in dynamic systems. *Hum. Fact.* 37, 1 (1995), 32–64.

Mica R. Endsley and Daniel J. Garland. 2000. *Situation Awareness Analysis and Measurement*. CRC Press.

Mica R. Endsley and Debra G. Jones. 2016. *Designing for Situation Awareness: An Approach to User-centered Design*. CRC Press.

Mica R. Endsley and Esin O. Kiris. 1995. The out-of-the-loop performance problem and level of control in automation. *Hum. Fact.* 37, 2 (1995), 381–394.

Larry P. English. 1999. *Improving Data Warehouse and Business Information Quality*. J. Wiley & Sons.

Martin J. Eppler and Peter Muenzenmayer. 2002. Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In *Proceedings of the International Conference on Information Quality (ICIQ'02)*. 187–196.

Joe Falco, Jeremy A. Marvel, and Richard Norcross. 2012. Collaborative robotics: Measuring blunt force impacts on humans. In *Proceedings of the International Conference on Safety and Industrial Automated Systems*.

P. D. Falorsi, S. Pallara, A. Pavone, A. Alessandroni, E. Massella, and M. Scannapieco. 2003. Improving the quality of toponymic data in the Italian public administration. In *Proceedings of the International Conference on Database Theory*, Vol. 3.

S. Alireza Fayazi, Nianfeng Wan, Stephen Lucich, Ardalan Vahidi, and Gregory Mocko. 2013. Optimal pacing in a cycling time-trial considering cyclist's fatigue dynamics. In *Proceedings of the American Control Conference*. IEEE, 6442–6447.

Fanny Ficuciello, Luigi Villani, and Bruno Siciliano. 2015. Variable impedance control of redundant manipulators for intuitive human–robot physical interaction. *IEEE Trans. Rob.* 31, 4 (2015), 850–863.

Paul M. Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *J. Exper. Psychol.* 47, 6 (1954), 381.

Stefan Flixeder, Tobias Glück, and Andreas Kugi. 2016. Modeling and force control for the collaborative manipulation of deformable strip-like materials. *IFAC-PapersOnLine* 49, 21 (2016), 95–102.

Terrence Fong, Clayton Kunz, Laura M. Hiatt, and Magda Bugajska. 2006. The human-robot interaction operating system. In *Proceedings of the ACM SIGCHI/SIGART Conference on Human-Robot Interactions*. ACM, 41–48.

Terrence Fong, Charles Thorpe, and Charles Baur. 2003. Collaboration, dialogue, human-robot interaction. In *Robotics Research*. Springer, 255–266.

Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *Proceedings of the International Symposium on Collaborative Technologies and Systems.* IEEE, 106–114.

Masakazu Fujii, Hiroki Murakami, and Mitsuharu Sonehara. 2016. Study on application of a human-robot collaborative system using hand-guiding in a production line. *IHI Eng. Rev* 49, 1 (2016), 24–29.

Ernesto Gambao, Miguel Hernando, and Dragoljub Surdilovic. 2012. A new generation of collaborative robots for material handling. In *Proceedings of the International Symposium on Automation and Robotics Construction*, Vol. 29. Vilnius Gediminas Technical University, Department of Construction Economics, 1.

Bernie Garrett, Tarnia Taverner, Diane Gromala, Gordon Tao, Elliott Cordingley, and Crystal Sun. 2018. Virtual reality clinical research: Promises and challenges. *JMIR Ser. Games* 6, 4 (2018), e10839.

Philip Godfrey. 2002. Overall equipment effectiveness. *Manuf. Eng.* 81, 3 (2002), 109–112.

K. Gold. 2009. An information pipeline model of human-robot interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interactions.* ACM, 85–92.

M. A. Goodrich and D. R. Olsen. 2003. Seven principles of efficient human robot interaction. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics.* IEEE, 3942–3948.

V. Groom and C. Nass. 2007. Can robots be teammates? Benchmarks in human-robot teams. *Interact. Stud., Psychol. Benchm. Hum.-rob. Interact.* 8 (2007), 483–500.

Andrea Guazzini, Alessandro Cini, Rosapia Lauro Grotto, and Franco Bagnoli. 2012. Virtual small group dynamics: A quantitative experimental framework. *Rev. Psychol. Front* 1 (2012), 10–17.

Edward T. Hall. 1963. A system for the notation of proxemic behavior. *Amer. Anthropol.* 65, 5 (1963), 1003–1026.

Sandra G. Hart. 2006. NASA-task load index (NASA-TLX): 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908.

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Adv. Psychol.* 52 (1988), 139–183.

Gerald Heddy, Umer Huzaifa, Peter Beling, Yacov Haimes, Jeremy Marvel, Brian Weiss, and Amy LaViers. 2015. Linear temporal logic (LTL) based monitoring of smart manufacturing systems. In *Proceedings of the Prognostic Health Management Society Conference.*

Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2010. Assessing acceptance of assistive social agent technology by older adults: The Almere model. *Int. J. Soc. Rob.* 2, 4 (2010), 361–375.

E. Helms, R. D. Schraft, and M. Hägele. 2002. rob@work: Robot assistant in industrial environments. In *Proceedings of the 11th IEEE International Workshop on Robot and Human Communications.* 399–404.

William E. Hick. 1952. On the rate of gain of information. *Quart. J. Exper. Psychol.* 4, 1 (1952), 11–26.

Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Trans. Hum.-Mach. Syst.* 49, 3 (2019).

R. Ishikawa, K. Sasaki, and I. Fukui. 1992. Standardization of Japanese version of FNE and SADS. *Koudou Ryouhou Kenkyu (Jap. J. Behav. Ther.)* 18 (1992), 10–17.

ISO. 1998. *ISO 9283 - Manipulating Industrial Robots - Performance Criteria and Related Test Methods.* Standard. International Organization for Standardization.

ISO. 2001. *ISO 9126-1 - Software Engineering - Product Quality - Part 1: Quality Model.* Standard. International Organization for Standardization.

ISO. 2010. *ISO 9241-210 - Ergonomics of Human-system Interaction - Part 210 - Human-centered Design for Interactive Systems.* Standard. International Organization for Standardization.

ISO. 2011a. *10218-2 - Robots and Robotic Devices - Safety Requirements - Part 2: Industrial Robot Systems and Integration.* Standard. International Organization for Standardization.

ISO. 2011b. *ISO 10218-1 - Robots and Robotic Devices - Safety Requirements - Part 1: Robots.* Standard. International Organization for Standardization.

ISO. 2011c. *ISO 25010 - Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - System and Software Quality Models.* Standard. International Organization for Standardization.

ISO. 2012. *ISO 8373 - Robotics and Robotic Devices - Vocabulary.* Standard. International Organization for Standardization.

ISO. 2016. *ISO/TS 15066 - Robotics and Robotic Devices - Collaborative Robots.* Standard. International Organization for Standardization.

Lars Christian Jensen, Kerstin Fischer, Franziska Kirstein, Dadhichi Shukla, Özgur Erkennt, and Justus Piater. 2017. It gets worse before it gets better: Timing of instructions in close human-robot collaboration. In *Proceedings of the International Conference on Human-Robot Interactions.* ACM, 145–146.

Manfred A. Jeusfeld, Christoph Quix, and Matthias Jarke. 1998. Design and analysis of quality information for data warehouses. In *Proceedings of the International Conference on Conceptual Modelling.* Springer, 349–362.

J. Johnson. 2014. *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines* (2nd ed.). Morgan Kaufmann.

Debra G. Jones. 2000. Subjective measures of situation awareness. In *Situation Awareness Analysis and Measurement.* Mahway, New Jersey: Lawrence Erlbaum Associates, 113–128.

Debra G. Jones and Mica R. Endsley. 1996. Sources of situation awareness errors in aviation.*Aviation, Space, Environ. Med.* 67, 6 (1996).

Michiel Joosse, Aziez Sardar, Manja Lohse, and Vanessa Evers. 2013. BEHAVE-II: The revised set of measures to assess users' attitudinal and behavioral responses to a social robot. *Int. J. Soc. Rob.* 5, 3 (2013), 379–388.

David B. Kaber and Mica R. Endsley. 2004. The effects of level of automation and adaptive automation on human performance, situation awareness, and workload in a dynamic control task. *Theor. Issues Erg. Sci.* 5, 2 (2004), 113–153.

Krishnanand N. Kaipa, Akshaya S. Kankanhalli-Nagendra, Nithyananda B. Kumbla, Shaurya Shriyam, Srudeep Somnaath Thevendria-Karthic, Jeremy A. Marvel, and Satyandra K. Gupta. 2016. Addressing perception uncertainty induced failure modes in robotic bin-picking. *Rob. Comput.-Integ. Manuf.* 42 (2016), 17–38.

Tavo Kangru, Jüri Riives, Tauno Otto, Meelis Pohlak, and Kashif Mahmood. 2018. Intelligent decision making approach for performance evaluation of a robot-based manufacturing cell. In *Proceedings of the International Mechanical Engineering Congress Exposition (ASME'18).* American Society of Mechanical Engineers, V002T02A092–V002T02A092.

Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *Int. J. Aviation Psych.* 3, 3 (1993), 203–220.

Michael Klug and James P. Bagrow. 2016. Understanding the group dynamics and success of teams. *Roy. Soc. Open Sci.* 3, 4 (2016), 160007.

Shirlee-Ann Knight and Janice Burn. 2005. Developing a framework for assessing information quality on the World Wide Web.*Informing Sci.* 8 (2005).

Yang W. Lee, Diane M. Strong, Beverly K. Kahn, and Richard Y. Wang. 2002. AIMQ: A methodology for information quality assessment. *Info. & Manag.* 40, 2 (2002), 133–146.

Michael Leonard, Suzanne Graham, and Doug Bonacum. 2004. The human factor: The critical importance of effective teamwork and communication in providing safe care. *Qual. Safety Health Care* 13, suppl 1 (2004), i85–i90.

Carl-Fredrik Lindberg, SieTing Tan, JinYue Yan, and Fredrik Starfelt. 2015. Key performance indicators improve industrial performance. *Energy Proc.* 75 (2015), 1785–1790.

Honghai Liu and George M. Coghill. 2005. A model-based approach to robot fault diagnosis. *Knowl.-based Syst.* 18, 4 (2005), 225–233.

David Loshin. 2001. *Enterprise Knowledge Management: The Data Quality Approach.* Morgan Kaufmann.

Lanssie Mingyue Ma, Terrence Fong, Mark J. Micire, Yun Kyung Kim, and Karen Feigh. 2018. Human-robot teaming: Concepts and components for design. In *Field and Service Robotics.* Springer, 649–663.

Keith R. MacArthur, Kimberly Stowers, and P. A. Hancock. 2017. Human-robot interaction: Proximity and speed—slowly back away from the robot! In *Advances in Human Factors in Robots and Unmanned Systems.*Springer, 365–374.

Jeremy Marvel, Elena Messina, Brian Antonishek, Lisa Fronczek, and Karl Van Wyk. 2015b. *NISTIR 8093: Tools for Collaborative Robots within SME Workcells.* Technical Report. National Institute of Standards and Technology.

Jeremy A. Marvel. 2013. Performance metrics of speed and separation monitoring in shared workspaces. *IEEE Trans. Autom. Sci. Eng.* 10, 2 (2013), 405–414.

Jeremy A. Marvel. 2014a. Collaborative Robotics: A Gateway into Factory Automation. Retrieved from http://news.thomasnet.com/imt/2014/09/03/collaborative-robots-a-gateway-into-factory-automation.

Jeremy A. Marvel. 2014b. Performance of Collaborative Robot Systems. Retrieved from https://www.nist.gov/programs-projects/performance-collaborative-robot-systems.

Jeremy A. Marvel, Roger Bostelman, and Joe Falco. 2018. Multi-robot assembly strategies and metrics. *ASM Comput. Surv.* 51, 1 (2018), 14:1–14:32.

Jeremy A. Marvel, Joe Falco, and Ilari Marstio. 2015a. Characterizing task-based human–robot collaboration safety in manufacturing. *IEEE Trans. Syst., Man, Cybern.: Syst.* 45, 2 (2015), 260–275.

Jeremy A. Marvel and Rick Norcross. 2017. Implementing speed and separation monitoring in collaborative robot workcells. *Rob. Comput.-Integ. Manuf.* 44 (2017), 144–155.

Björn Matthias, Susanne Oberer-Treitz, and Hao Ding. 2014. Experimental characterization of collaborative robot collisions. In *Proceedings of the 41st International Symposium on Robotics.* VDE, 1–6.

James C. McCroskey, Michael D. Scott, and Thomas J. Young. 1973. Measurement of the credibility of peers and spouses. In *Proc. Ann. Conf. Int. Comm. Assoc.* ERIC.

Everett N. McKay. 2013. *UI is Communication: How to Design Intuitive, User Centered Interfaces by Focusing on Effective Communication.* Newnes.

George Michalos, Niki Kousi, Panagiotis Karagiannis, Christos Gkournelos, Konstantinos Dimoulas, Spyridon Koukas, Konstantinos Mparis, Apostolis Papavasileiou, and Sotiris Makris. 2018. Seamless human robot collaborative assembly–An automotive case study. *Mechatronics* 55 (2018), 194–211.

Mark Micire, Munjal Desai, Jill L. Drury, Eric McCann, Adam Norton, Katherine M. Tsui, and Holly A. Yanco. 2011. Design and validation of two-handed multi-touch tabletop controllers for robot teleoperation. In *Proceedings of the 16th International Conference on Intelligent User Interfaces.* ACM, 145–154.

Holmes Miller. 1996. The multiple dimensions of information quality. *Info. Syst. Manag.* 13, 2 (1996), 79–82.

Bonnie M. Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.

R. R. Murphy and J. L. Burke. 2005. Up from the rubble: Lessons learned about HRI from search and rescue. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* 437–441.

R. R. Murphy, D. Riddle, and E. Rasmussen. 2004. Robot-assisted medical reachback: A survey of how medical personnel expect to interact with rescue robots. In *Proceedings of the IEEE International Workshop on Robot and Human Communications.* IEEE, 301–306.

National Instruments. 2016. LabVIEW for Industrial Robotics. Retrieved from http://www.ni.com/robotics/industrial/.

Takumi Ninomiya, Akihito Fujita, Daisuke Suzuki, and Hiroyuki Umemuro. 2015. Development of the multi-dimensional robot attitude scale: Constructs of people's attitudes towards domestic robots. In *Proceedings of the International Conference on Social Robotics.* Springer, 482–491.

Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interact. Stud.* 7, 3 (2006), 437–454.

Brid O'Conaill and David Frohlich. 1995. Timespace in the workplace: Dealing with interruptions. In *Proceedings of the Conference on Human Factors in Computer Systems.* ACM, 262–263.

D. R. Olsen and M. A. Goodrich. 2003. Metrics for evaluating human-robot interactions. In *Proceedings of the NIST Performance Metrics of Intelligent Systems Workshop.* NIST.

M. Orcutt. 2014. Robots Rising. Retrieved from https://www.technologyreview.com/s/529971/robots-rising/.

Gergely Palla, Albert-László Barabási, and Tamás Vicsek. 2007. Quantitative social group dynamics on a large scale. *Communities* 5 (2007), 23.

X. Perrin, R. Chavarriaga, C. Ray, R. Siegwart, and R. X. Millán. 2008. A comparative psychophysical and EEG study of different feedback modalities for HRI. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interactions.* ACM, 41–48.

Luka Peternel, Nikos Tsagarakis, Darwin Caldwell, and Arash Ajoudani. 2016. Adaptation of robot physical behaviour to human fatigue in human-robot co-manipulation. In *Proceedings of the IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids'16).* IEEE, 489–494.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. In *Proceedings of the National Academy of Sciences*, Vol. 108. National Academy of Sciences, 3526–3529.

Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Commun. ACM* 45, 4 (2002), 211–218.

Frederick Proctor, Stephen Balakirsky, Zeid Kootbally, Thomas Kramer, Craig Schlenoff, et al. 2016. The canonical robot command language (CRCL). *Industr. Rob.: Int. J.* 43, 5 (2016), 495–502.

Lisa Rebenitsch and Charles Owen. 2016. Review on cybersickness in applications and visual displays. *Virt. Real.* 20, 2 (2016), 101–125.

Gary B. Reid and Thomas E. Nygren. 1988. The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Adv. Psychol.* 52 (1988), 185–218.

Julie Rennecker and Lindsey Godwin. 2005. Delays and interruptions: A self-perpetuating paradox of communication technology use. *Info. Organiz.* 15, 3 (2005), 247–266.

D. R. Riddle, R. R. Murphy, and J. L. Burke. 2005. Robot-assisted medical reachback: Using shared visual information. In *Proceedings of the IEEE International Workshop on Robot and Human Communications.* IEEE, 635–642.

Teresa L. Roberts and Thomas P. Moran. 1983. The evaluation of text editors: Methodology and empirical results.*Commun. ACM* 26, 4 (1983), 265–283.

G. Rodriguez and C. Weisbin. 2003. A new method to evaluate human-robot system performance. *Auton. Rob.* 14, 2 (2003), 165–178.

Susana Rubio, Eva Díaz, Jesús Martín, and José M. Puente. 2004. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Appl. Psychol.* 53, 1 (2004), 61–86.

Behzad Sadrfaridpour, Hamed Saeidi, Jenny Burke, Kapil Madathil, and Yue Wang. 2016. Modeling and control of trust in human-robot collaborative manufacturing. In *Robust Intelligence and Trust in Autonomous Systems.*Springer, 115–141.

John J. Salerno, Erik P. Blasch, Michael Hinman, and Douglas M. Boulware. 2005. Evaluating algorithmic techniques in supporting situation awareness. In *Proceedings of the Defense and Security Conference.* 96–104.

Monica Scannapieco, Antonino Virgillito, Carlo Marchetti, Massimo Mecella, and Roberto Baldoni. 2004. The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. *Info. Syst.* 29, 7 (2004), 551–582.

J. Scholtz, J. Young, and J. L. Drury. 2004. Evaluation of human-robot interaction awareness in search and rescue. In *Proceedings of the IEEE International Conference on Robotics and Automation.* IEEE, 2327–2332.

William Shackleford, Geraldine Cheok, Tsai Hong, Kamel Saidi, and Michael Shneier. 2016. Performance evaluation of human detection systems for robot safety. *J. Intell. Rob. Syst.* 83, 1 (2016), 85–103.

Julie Shah and Cynthia Breazeal. 2010. An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming. *Hum. Fact.* 52, 2 (2010), 234–245.

Jane Shi, Glenn Jimmerson, Tom Pearson, and Roland Menassa. 2012. Levels of human and robot collaboration for automotive manufacturing. In *Proceedings of the Workshop on Performance Metrics of Intelligent Systems.* ACM, 95–100.

S. Shibata. 2016. Development of the feelings toward nature scale and relationship between feelings toward nature and proximity to nature. *Shinrigaku Kenkyu (Jap. J. Psych.)* 87, 1 (2016), 50–59.

B. Shneiderman. 1998. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison-Wesley.

Michael O. Shneier, Elena R. Messina, Craig I. Schlenoff, Frederick M. Proctor, Thomas R. Kramer, and Joseph A. Falco. 2015. *NISTIR 8090: Measuring and Representing the Performance of Manufacturing Assembly Robots.* Technical Report. National Institute of Standards and Technology.

Daniel J. Simons and Christopher F. Chabris. 1999. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* 28, 9 (1999), 1059–1074.

Daniel J. Simons and Daniel T. Levin. 1997. Change blindness. *Trends Cogn. Sci.* 1, 7 (1997), 261–267.

George S. Snoddy. 1926. Learning and stability: A psychophysiological analysis of a case of motor learning with clinical applications. *J. Appl. Psychol.* 10, 1 (1926), 1.

David A. Sousa. 2006. *How the Brain Learns.* SAGE Publications Ltd.

Neville A. Stanton, Peter R. G. Chambers, and J. Piggott. 2001. Situational awareness and safety. *Safety Sci.* 39, 3 (2001), 189–204.

Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interactions.* 33–40.

Ying Su and Zhanming Jin. 2006. A methodology for information quality assessment in the designing and manufacturing process of mechanical products. In *Information Quality Management: Theory and Applications.* Idea Group Publishing Hershey, PA, USA, 190–220.

Nassim Nicholas Taleb. 2007. *The Black Swan: The Impact of the Highly Improbable.* Vol. 2. Random House.

Adriana Tapus and Maja J. Mataric. 2008. Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *Proceedings of the AAAI Spring Symposium: Emotion, Personality, and Social Behavior.* 133–140.

R. M. Taylor. 1990. Situational awareness rating technique (SART): The development of a tool for aircrew systems design. *AGARD, Situat. Aware. Aero. Operat.* AGARD-CP-478 (1990), 3:1–3:17.

Douglas S. Thomas. 2018. *The Costs and Benefits of Advanced Maintenance in Manufacturing.* US Department of Commerce, National Institute of Standards and Technology.

Pamela S. Tsang and Velma L. Velazquez. 1996. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 39, 3 (1996), 358–381.

Daniel Ullman and Bertram F. Malle. 2018. What does it mean to trust a robot?: Steps toward a multidimensional measure of trust. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interactions.* 263–264.

S. Valenti, A. Cucchiarelli, and M. Panti. 2002. Computer based assessment systems evaluation via the ISO 9126 quality model. *J. Info. Tech. Educ.* 1 (2002), 157–175.

Athulan Vijayaraghavan, Will Sobel, Armando Fox, David Dornfeld, and Paul Warndorf. 2008. Improving machine tool interoperability using standardized interface protocols: MTConnect. In *Proceedings of the International Symposium on Flexible Automation.*

Valeria Villani, Fabio Pini, Francesco Leali, and Cristian Secchi. 2018. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics* 55 (2018), 248–266.

Wayne L. Waag and Michael R. Houck. 1994. Tools for assessing situational awareness in an operational fighter environment.*Aviat., Space, Envir. MD* 65, 5 (1994), A13–A19.

Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. 2012a. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization.*Psychol. Bull.* 138, 6 (2012), 1172.

Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R. Pomerantz, Peter A. van der Helm, and Cees van Leeuwen. 2012b. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations.*Psychol. Bull.* 138, 6 (2012), 1218.

Jing-Jy Wang. 2005. Psychological abuse behavior exhibited by caregivers in the care of the elderly and correlated factors in long-term care facilities in Taiwan. *J. Nurs. Res.* 13, 4 (2005), 271–280.

Richard Y. Wang. 1998. A product perspective on total data quality management. *Commun. ACM* 41, 2 (1998), 58–65.

Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *J. Manag. Info. Syst.* 12, 4 (1996), 5–33.

Astrid Weiss and Andreas Huber. 2016. User experience of a smart factory robot: Assembly line workers demand adaptive robots. *arXiv preprint arXiv:1606.03846* (2016).

Brian A. Weiss, Michael P. Brundage, Yannick Tamm, Tommi Makila, and Joan Pellegrino. 2019. *Summary Report on the Industry Forum for Monitoring, Diagnostics, and Prognostics for Manufacturing Operations.* Technical Report. National Institute of Standards and Technology.

Brian A. Weiss, John Horst, and Fred Proctor. 2013. *NISTIR 7911: Assessment of Real-time Factory Performance Through the Application of Multi-relationship Evaluation Design.* Technical Report. National Institute of Standards and Technology.

Brian A. Weiss and Linda C. Schmidt. 2010. The multi-relationship evaluation design framework: Creating evaluation blueprints to assess advanced and intelligent technologies. In *Proceedings of the Performance Metrics of Intelligent Systems Workshop.* NIST, 136–145.

Brian A. Weiss, Michael Sharp, and Alexander Klinger. 2018. Developing a hierarchical decomposition methodology to increase manufacturing process and equipment health awareness. *J. Manuf. Syst.* 48 (2018), 96–107.

Katherine S. Welfare, Matthew R. Hallowell, Julie A. Shah, and Laurel D. Riek. 2019. Consider the human work experience when integrating robotics in the workplace. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interactions.* IEEE, 75–84.

C. D. Wickens, J. D. Lee, Y. Liu, and S. Gordon-Becker. 2004. *An Introduction to Human Factors Engineering.* Pearson Prentice Hall.

L. Wood. 2014. Research and Markets: Global Articulated Robots Market Growth of 16.27% CAGR by 2019—Analysis, Technologies & Forecasts 2016–2019. Retrieved from http://www.businesswire.com/news/home/20160108005449/en/Research-Markets-Global-Articulated-Robots-Market-Growth.

Andrea Maria Zanchettin, Nicola Maria Ceriani, Paolo Rocco, Hao Ding, and Björn Matthias. 2016. Safety in human-robot collaborative manufacturing environments: Metrics and control. *IEEE Trans. Autom. Sci. Eng.* 13, 2 (2016), 882–893.

L. Zhao, X. Lu, and Y. Hu. 2018. A proposed theoretical model of discontinuous usage of voice-activated intelligent personal assistants (IPAs) a proposed theoretical model of discontinuous. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS'18).*

Li Zhu, Jacob Meivik, Charlotta Johnsson, Kristofer Bengtsson, Hakan Pettersson, Martina Varisco, and Massimiliano M. Schiraldi. 2018. Key performance indicators in manufacturing operations management: A case study of the IS022400-standard applied at Volvo cars. In *Proceedings of the IEEE International Conference on Emerging Technologies in Factory Automation (ETFA'18)*, Vol. 1. IEEE, 1149–1152.

Received August 2018; revised February 2020; accepted February 2020