Safety Assessment of Collaborative Robotics through Automated Formal Verification

Federico Vicentini, Mehrnoosh Askarpour, Matteo G. Rossi, and Dino Mandrioli

***Abstract*—A crucial aspect of physical human-robot collaboration (HRC) is to maintain a safe common workspace for human operator. However, close proximity between human-robot and unpredictability of human behavior raises serious challenges in terms of safety. This paper proposes a risk analysis methodology for collaborative robotic applications which is compatible with well-known standards in the area and relies on formal verification techniques to automate the traditional risk analysis methods. In particular, the methodology relies on temporal logic-based models to describe the different possible ways in which tasks can be carried out, and on fully automated formal verification techniques to explore the corresponding state-space to detect and modify the hazardous situations at early stages of system design.**

***Index Terms*—Formal Methods, Temporal Logic, Robot Safety, Model-based Risk Assessment, Human-Robot Collaboration**

I. Introduction

Human-robot collaboration (HRC) in industrial settings enhances the flexibility and adaptability of robotic systems for production. Close proximity and hybrid task assignment between humans and robots have substantial impacts on the safety of human operators. Most importantly, shared dynamic environments cannot be effectively supported by static safety analyses. In particular, the hybrid human-robot tasks generate different possible execution workflows. Therefore, various task assignments among humans and robots can also change during operations. In realistic scenarios, the environment can also change, due to mobile resources, tool changing, modification of the layout and process locations, so that the specifications of the system need to be re-evaluated. The safety of such evolving systems should not be verified as a static property, but rather according to the behavior of the system in actual situations.

A HRC safety analysis shall comply with the requirements of ISO 12100 [1], as shown in the flowchart at the top of fig. 1. First, the procedure investigates any source of potential harm or discomfort to humans—known as *hazards*—during the course of a task execution. Then, identified hazards are evaluated in terms of *risk*—a metric, combining the intensity and likelihood of hazards, for the potential hazard to become an actual harm—and mitigated by introducing suitable *risk reduction measures (RRM)*. A hazard is just a potential danger, and the goal of safety is achieved when *high risks* are resolved using different strategies [2].

For changing scenarios, safety analysis needs to be comprehensive and exhaustive, and it should facilitate the exploration of different variants and configurations of the system (e.g., to support design updates). Manual risk assessments applied to HRC cases are necessarily static, and pose significant challenges to experts in identifying all possible hazardous situations due to the high level of unpredictability of typical human behaviors. As applications become more and more complex and human-robot interactions expand, some risks could be overlooked because they are hidden in the evolution of the system. The main resulting limitation is the likely selection of ineffective or over-conservative strategies (e.g., *always* keep a low speed).

This paper introduces a rigorous, iterative methodology—based on formal methods—to guarantee the safety of physical HRC systems. It is structured as follows: Section II discusses the state of the art in the field of safety analysis of HRC applications, and Section III gives an overview of our own contributions to it. Section IV introduces the background in formal verification relevant for safety assessment, highlighting models and safety properties. HRC system modeling is described in Section V, which introduces the formalization of all relevant elements in terms of temporal logic, including the modeling of actions which is used to identify safety-critical situations. Section VI describes how the formal approach supports the key modules for safety analysis: extraction of hazards from the formal model, risk estimation, and the verification of risk reduction upon the introduction of models for mitigation measures. Section VII reports on a use case of collaborative task where the method is applied and illustrated. Finally, Section VIII discusses several aspects of the methodology, the obtained results and some open issues.

II. State of the art

There is a rich literature about safety analysis of systems where humans and machines interact in a critical way. We can coarsely categorize it as: (a) general, mostly informal, safety analysis techniques; (b) techniques and standards that more specifically address HRC systems; (c) methods, techniques and tools that exploit—at some level—the mathematical formalisms. Unsurprisingly, some works found in literature exhibit features that belong to several of the above categories.

Among traditional, general-purpose, safety-centered techniques, FMEA [3], FMECA [4], FTA [5], HAZOP [6] and STAMP [7] are very well known. FMEA is a qualitative approach that detects the potential failures of a system, evaluates their effects, introduces actions to eliminate or reduce the chance of the potential failure and documents the whole process. FMECA extends FMEA with criticality analysis and quantifies effects and severities of failures. Given an identified failure, these two methods basically find its main origin and offer detailed intuitions about the system. FTA is a deductive

failure analysis technique which exploits boolean algebra and probability theory. It finds out the causes of failure behavior via diagram and logic models [8]. HAZOP (HAZard and OPerability study, [6]) identifies hazards and operability problems and investigates how a system can deviate from the intended design. HAZOP is usually carried out when the detailed design has been completed. Then, a multi-disciplinary team (HAZOP Team) uses brainstorming meetings to determine if the design intents are satisfied, using guide-words that provide qualitative measures of design deviations.

Unlike other methods, STAMP (Systems-Theoretic Accident Model and Processes, [7]) analyzes system accidents as a result of inadequate control (e.g., external events—including human factors—failures in the interaction among subsystems or between human and system) and not simply as a chain of failures. STPA (STAMP-Based Process Analysis, [9]) is a powerful hazard analysis technique based on STAMP which is used in different industry sectors.

None of these methods, however, specifically addresses the requirements of the safety analysis of HRC applications and their peculiarities. These are tackled by the works of category (b) above, of which the following are some of the most representative ones: [10] deals with unpredictable environments and the various temporal combinations of their evolution, capturing hazards due to human factors; [11] focuses on the combination of different hazards, avoiding to create an excessive amount of redundant information; [12] argues for producing results that are not generic, (semi)-automated and reusable for different scenarios.

Formal methods [13]—those of type (c) in the classification above—are a set of well-defined mathematical languages and techniques for accurate modeling, specification, and automated verification of systems. The formal model of an application can capture all of its associated workflows. Additionally, verification can be done via automated techniques that are capable of exploring all such workflows, thus eliminating any chance of overlooking any of them.

In robotics, formal modeling and verification techniques have been mostly used for defining robot behavior (e.g., controllers, motion planning, fault avoidance); therefore, safety has been tackled only indirectly.

Formal verification has been also exploited for exhaustive and comprehensive safety analysis. A number of works developed methods that employ (semi-)formal[1] approaches to address safety. For example, [14], [15], [16] merge the semi-formal UML notation [17] with HAZOP first to describe test-case scenarios through Sequence and State Machine Diagrams, and then to identify hazards and analyze their risks. [18] use influence diagrams to model and represent tasks and their interrelations, and introduce a task-learning technique that allows human operators to teach robots the tasks to be executed. [19], [20] model an assistant robotic system through the Brahms language [21] and verify it against safety requirements via the SPIN model checker [22]. [23], [24] also employ formal verification to analyze the safety of assistant robots. [25], [26], [27], instead, focus on non-robotic collaborative applications and use simplified cognitive architectures that require large manual customizations, but mostly ignore human fallibility and plausible errors.

The aforementioned works, however, take into limited consideration the physical interaction between humans and robots, focusing on assistive robots and avoiding relevant mechanical hazards—like contacts—in industrial setups, as referenced in domain-specific international standards such as ISO 15066 and ISO 10218-2 [28] and [29].

To this end, task-centric approaches usually rely on decompositions of the tasks to be executed. For instance, Hierarchical Task Analysis (HTA [30]) offers an exhaustive description of tasks by decomposing them in a hierarchy of goals, sub-goals, operations and plans. Many task-centric approaches, e.g., [31], rely on HTA to analyze the safety hazards that the task itself might create. Models, however, should describe not only the task to be executed with its goals and workflow, but also, as pointed out by [32], the various components of the system: physical and functional elements of the work domain, human-machine interaction, as in [33], human judgments in different situations, the configuration of the control system and so on. This entails the necessity to create a well-integrated model of the overall system model, where human behavior plays a key role with distinguishing features. [34] identifies four categories of techniques for the modeling of the human behavior (many of which based on formal methods):

- *Human-device interface (HDI)* models, extensively studied in [35], focus on the interaction via interface tools and do not deal with the *co-presence* of human and device [36].
- *Cognitive* models describe the knowledge behind human observable behavior [35] and embodies a wide range of models, from classic (e.g., SOAR [37], ACT-R [38], [39]) to more recent ones (e.g., OCM [40], [41], PUM [42]). PUM is the best-suited for formal verification because of its general and accurate semantics. Examples of PUM-based techniques are [43], [44], [26], [45].
- *Human mental* models distinguish the machine model from the user's perception and cognition. [46], [25], [47], [48] introduce several formalisms for building such models. These models are too detailed to suit an exhaustive state-space and risk exploration because of state-explosion issues.
- *Task-analytic* models are based on hierarchical structures of tasks [34]; examples are OFM [49], UAN [50], ConcurTaskTrees [51] and EOFM [52], [53].

[27], [54], [55], [56], [57] combine cognitive models and probability distributions. However, the authors of these works mention that, since real data on human behavior is not available, they had to rely on self-generated fictional data that does not necessarily reflect or abstract the actual behavior of operators. This significantly lessens the reliability of these approaches for what concerns safety analysis.

Unlike interface models, cognitive and task-analytic ones are able to represent the collaboration between human operators and robots. However, they have been mostly used in contexts that are not fully formal, or focus on a limited set of

[1]The term "semi-formal" denotes a notation whose syntax—i.e., the rules for text or diagram composition—is mathematically defined, but whose semantics is left to the experience and intuition of writers and readers.

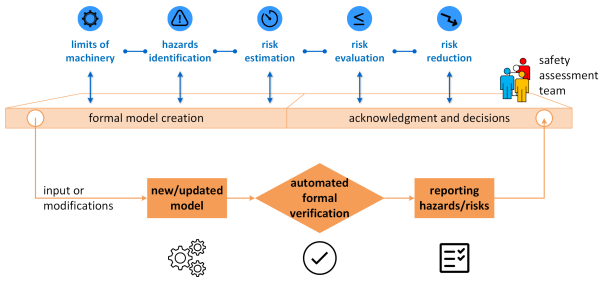


Figure 1: Overview of the SAFER-HRC methodology (bottom part) with respect to risk assessment steps, simplified from ISO 12100 (top part). SAFER-HRC creates a formal model of the HRC application, on which a risk assessment is carried out through automated formal verification. The results are reported to a team of human assessors, who acknowledge risk evaluations and refine the design of risk reduction models, then re-run SAFER-HRC until a desired result is achieved.

intended behaviors (i.e., regular use of the system, standard and intended instructions of the application), ignoring instead the *unintended* ones (i.e., any deviation from the intended use), which are a major source of hard-to-predict hazards [58]. HRC applications require a more comprehensive effort in the modeling of the human behavior than what is currently used in the domains of HDI—where co-presence is not an issue—or Human-Robot Interaction (HRI)—where human and robot co-exist in a shared environment, but no collaboration is required and the human is not an actor like the robot (e.g., service robotics, caregiver robots).

III. Contributions

This paper proposes a methodology called SAFER-HRC (Safety Assessment through Formal Verification in HRC) for supporting dynamic safety assessment in HRC applications. SAFER-HRC uses model-based formal verification to (i) exhaustively explore all possible execution workflows of an application, (ii) automatically identify hazards within those workflows and determine their associated risk, in compliance with ISO 12100 [1] and ISO 10218-2 [29], and (iii) introduce provisions for risky hazards and iteratively verify the effectiveness of such protective measures by computing residual risks.

The system's model is formalized in terms of temporal logic and is formally verified [59] against a desired safety property, i.e., the *negligibility of risks* (a hazard is harmless if its associated risk score is below a certain threshold). The formal verification is performed via the Zot [60] tool.

In our approach, the informal and goal-oriented description of an application is translated into a logic model that captures the key components of a HRC system (operators, robots and layout), complemented with formalized descriptions of hazardous situations. Additional modules formalize risk estimation procedures (as in ISO/TR 14121-2 [61]) and protective measures in the same logic. These modules include predicates and formulae about humans, robots, environment, hardware properties, spatial and functional relationships among elements, and safety protection factors (e.g., power limits). The whole logical model is *discretized* to abstract away from excessive details and to avoid state-space explosion issues [62].

The unpredictability of human behavior during the execution of a HRC task could generate a range of unintended, potentially hazardous, situations. Thus, a sufficiently accurate model of the system requires also to consider human factors and the observable manifestation of the operator's physical presence and activities. SAFER-HRC captures the effects of human behavior on safety properties to be verified, rather than cognitive-based mechanisms behind such behaviors. A functional model—rather than a cognitive one—replicates human errors independently from attitudes of operators (e.g., being tired or absent-minded) leading to such errors.

The selection of applicable risk mitigation strategies is not fully automated. As shown in fig. 1, our methodology does not replace the human risk assessor, but provides an automated assistant that formalizes the standard risk evaluation procedure of ISO 12100 applied to collaborative robotics. The human risk assessor is in charge of (i) tailoring the model for the system, and (ii) confirming the chosen mitigation action (e.g., limit power, reduce speed, change path) after running the verification process again.

IV. Background: Temporal Logic and Automated Formal Verification

The ultimate goal of the presented methodology is to comprehensively explore and verify all possible interactions between robots and human operators, with respect to *risks*. Recall that risks, not hazards, are the key notion of safety. The goal is then achieved by combining two elements: (i) a formal model that captures the sequences of actions performed by robots and operators in their fulfillment of assigned tasks; (ii) mechanisms to exhaustively analyze these sequences of actions, and to detect those that are deemed unacceptable if some hazards, when present, display an intolerable risk.

Modeling actions along time

The first requirement (modeling actions) is accomplished using a temporal logic language, called TRIO [63], which is capable of expressing the evolution over time of phenomena of interest (e.g., "the operator enters a robotic cell *before* a robot moves"). Additionally, the TRIO language features a quantitative notion of time, so that it is possible to render properties of tasks such as "the robot *always* reaches the working pose 3 time units *after* leaving its homing position". We rely on a fully logic-based approach to model actions because this allows us great flexibility and modularity in building models. For example, adding or removing actions to/from workflows simply involves adding/removing the corresponding logic formulae, and similarly for other elements of the system, such as RRMs. We chose TRIO as logic formalism for its expressiveness and amenability to automated formal verification (through Zot). Other logic languages, such as MTL [64], which can also be automatically analyzed through the Zot tool [65], would have suited our needs, though TRIO was developed specifically with the idea of allowing practitioners to build formal specifications of critical systems [66].

Table I: List of derived TRIO operators; ϕ, ψ denote propositions, and v is a variable and d is a constant value.

TRIO Operator	Definition	Meaning
$\mathsf{Past}(\phi, d)$	$d > 0 \wedge \mathsf{Dist}(\phi, -d)$	ϕ occurred d time units in the past
$\mathsf{Futr}(\phi, d)$	$d > 0 \wedge \mathsf{Dist}(\phi, d)$	ϕ occurs d time units in the future
$\mathsf{Alw}(\phi)$	$\forall t(\mathsf{Dist}(\phi, t))$	ϕ always holds
$\mathsf{Som}(\phi)$	$\exists t(\mathsf{Dist}(\phi, t))$	ϕ occurs sometimes
$\mathsf{Until}(\phi, \psi)$	$\exists t(\mathsf{Futr}(\psi, t) \wedge (\forall t'(0 < t' < t) \Rightarrow \mathsf{Dist}(\phi, t')))$	ψ will eventually occur and ϕ will hold till then
$\mathsf{Until_w}(\phi, \psi)$	$\mathsf{Until}(\phi, \psi) \vee \mathsf{Alw}(\phi)$	weak until: ψ may never occur in the future
$\mathsf{WithinF}(\phi, d)$	$\exists t(0 < t < d \wedge \mathsf{Dist}(\phi, t))$	ϕ will occur within d time units

TRIO formulae are built out of propositions and predicates describing the basic phenomena of the system (e.g., turn_on, off) and the usual propositional connectives—"and" ($\wedge$), "not" ($\neg$), etc.—and first-order quantifiers—"forall" ($\forall$), "exists" ($\exists$)—as well as a single basic modal operator, called Dist, that relates the *current time* to another time instant. Then, if ϕ is a TRIO formula and d is a time distance, $\mathsf{Dist}(\phi, d)$ holds at time t if, and only if, ϕ holds at time $t + d$. While TRIO can exploit both discrete and dense sets as time domains (e.g., the set of nonnegative integers $\mathbb{N}$, or the set of nonnegative real numbers $\mathbb{R}_{\geqslant 0}$), in this work we assume $\mathbb{N}$ as discrete time domain. For convenience in the writing of specification formulae, TRIO defines a number of *derived* temporal operators from the basic Dist, through propositional composition and first-order logic quantification. Table I reports some of the most significant ones, including those used in this work. For example, consider the following two generic predicates: leave_home, to describe when the robot departs from a homing position—i.e., leave_home is true at time t if the robot departs at that time—and at_target, to capture the fact that the robot is operational in a position. Then, the situation informally introduced above is formally captured by the TRIO formula $\mathsf{Alw}(\mathsf{leave\_home} \Rightarrow \mathsf{Dist}(\mathsf{at\_target}, 3))$. Interested readers can find more details about the TRIO language in [63].

A formula ϕ describing some situation in the TRIO language is evaluated over *histories*, which formally capture the notion of "behavior" of a system.

Definition 1: Given a predicate pred, a **history** H for pred is a set of its values along the temporal domain $\mathbb{N}$; we indicate the values taken by pred in history H as ${}^{H}\overline{\mathsf{pred}}(0), {}^{H}\overline{\mathsf{pred}}(1), {}^{H}\overline{\mathsf{pred}}(2), \ldots$. This definition is extended in a natural way to sets of predicates $\mathcal{S}$ and to formulae ϕ.

For example, consider the predicate leave_home introduced above: a history H_1 might be such that leave_home is true at time 8 (denoted as ${}^{H_1}\overline{\mathsf{leave\_home}}(8)$). Another possible history H_2 might be such that leave_home is true at time 10, but not at time 8.

Definition 2 (Semantics of Dist*):* Let ϕ be a TRIO formula, $\mathbb{N}$ the temporal domain and $t \in \mathbb{N}$ the current time instant. Then ${}^{H}\overline{\phi}(t)$ indicates the value of formula ϕ at instant t for history H, which is either *true* or *false*. Let d be a (positive or negative) time distance—i.e., $d \in \mathbb{Z}$. Formula $\mathsf{Dist}(\phi, d)$ holds at time t for history H if, and only if, ϕ holds at time $t + d$, or, equivalently, ${}^{H}\overline{\mathsf{Dist}(\phi, d)}(t)$ is true if, and only if, ${}^{H}\overline{\phi}(t + d)$ is true.

For example, consider history H_1 (see fig. 2) such that both ${}^{H_1}\overline{\mathsf{leave\_home}}(8) = true$ and ${}^{H_1}\overline{\mathsf{at\_target}}(11) = true$ hold; then, formula $\phi_{ex} := \mathsf{leave\_home} \Rightarrow \mathsf{Dist}(\mathsf{at\_target}, 3)$ holds at time 8 in H_1, that is, ${}^{H_1}\overline{\phi_{ex}}(8) = true$.

Verification of properties

A history H is said to *satisfy* a TRIO formula ϕ if ϕ holds for H in the origin of the time domain—i.e., if ${}^{H}\overline{\phi}(0)$ is true. In general, a formula ϕ is *satisfiable* if there exists a history H for which it is satisfiable; if no such history exists, ϕ is *unsatisfiable*. For example, consider formula $\mathsf{Alw}(\phi_{ex})$, where ϕ_{ex} was introduced above, and the histories H_1 and H_2 of fig. 2. History H_1 satisfies $\mathsf{Alw}(\phi_{ex})$, because every time leave_home is true in the history (i.e., at instants 1 and 8), after 3 time units at_target is also true; hence, ${}^{H_1}\overline{\mathsf{Alw}(\phi_{ex})}(0) = true$. History H_2, on the other hand, does not satisfy $\mathsf{Alw}(\phi_{ex})$, because leave_home is true at time 1, but there is no corresponding at_target at time 4, hence ϕ_{ex} is not true in 1 and ${}^{H_2}\overline{\mathsf{Alw}(\phi_{ex})}(0) = false$.

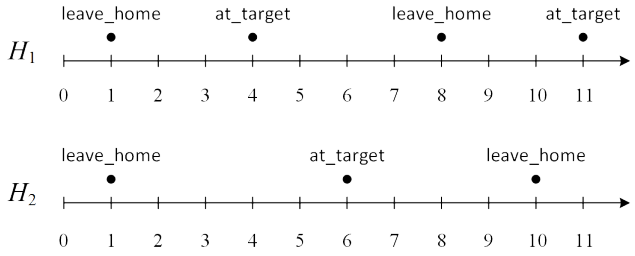


Figure 2: Two examples of histories for formula $\mathsf{Alw}(\mathsf{leave\_home} \Rightarrow \mathsf{Dist}(\mathsf{at\_target}, 3))$, one that satisfies the formula (history H_1) and one that does not satisfy it (H_2).

Satisfiability checking is the problem of determining whether a TRIO formula ϕ is satisfiable or not. In general, checking the satisfiability of TRIO formulae is an undecidable problem. However, this paper uses a decidable subset of the language, that can be handled by automated tools, in order to build the system model and to express its properties. In particular, the Zot bounded satisfiability checker [60] implements several techniques for checking the satisfiability of TRIO formulae in an automated manner [65], [67]. More precisely, given an input TRIO formula ϕ, Zot returns either a history satisfying ϕ or the value UNSAT, which stands for "unsatisfiable". Satisfiability checking is the underlying mechanism for verifying the safety properties of an application. An application corresponds to a set of histories. Let ϕ_S be a TRIO formula that formalizes the possible behaviors of a system. Then, the histories that satisfy ϕ_S are exactly those that capture

all possible sequences of phenomena of interest occurring in the application. Let ϕ_P be a TRIO formula that captures a safety requirement—e.g., "the risk value is always less then a given tolerable threshold". Then, the desired property to be verified holds for the application if there are no histories of the system described by ϕ_S that violate the property described by ϕ_P. Formally, there are no behaviors (histories) of the system for which "not ϕ_P" holds.

Definition 3 (Safety verification): Given a formula ϕ_S capturing the behaviors of the system and a formula ϕ_P formalizing the safety property, the system is safe if, and only if, $\phi_S \wedge \neg\phi_P$ is unsatisfiable.

Operationally, an application is verified to be safe or not by inputting a suitable TRIO formula (i.e., $\phi_S \wedge \neg\phi_P$) in the Zot bounded satisfiability checker. The outcome UNSAT proves the application being safe. Otherwise, Zot returns a history that not only satisfies ϕ_S, but also violates ϕ_P (because it satisfies $\neg\phi_P$); this is a *counterexample* witnessing why the application is not safe.

Definition 4: Under the generic notion of risk, the specific **safety property** ϕ_P to be used in definition 3 becomes

$$\mathsf{Alw}\,(\mathsf{risk} \leqslant \tau) \tag{1}$$

where τ is a tolerable value of a quantitative notion of risk (to be detailed in Subsection VI-B).

The description of an application is made of a set—or a conjunction—of formulae ϕ_S, to which definition 3 collectively applies. An instance of the formal model captures all possible behaviors of the system in a given application—i.e., the histories of values taken by temporal logic predicates. Different histories from the same model instance are possible and are due to different timings for the modeled actions (e.g., an operator might reach a target in a non-deterministic way). Whenever an application is changed, a new instance of the TRIO model is produced, in turn corresponding to many histories. Changes might be due to variations in the nominal workflow when the functional purpose of the application is modified, or to variations in the properties of elements (e.g., safety limitations are enabled in robots), a rearrangement of the layout, the introduction of new devices, and so on. Upon re-instantiation of the model—regardless of the reason—the safety condition captured by formula (1) is checked again for satisfiability. This automatic mechanism is very useful whenever the conditions of the application change because a planning action is done (e.g., task change) or some safety functions (e.g., power limitation) are manually enabled by the safety person(s) in charge of the tool.

V. Methodology of formalization of HRC applications

This section provides the modeling elements of human-robot collaborative tasks—i.e., the components of the system (subsection V-A)—and how such tasks are organized in actions (subsection V-B). The model described in this section and in section VI constitutes the temporal logic formulae denoted by ϕ_S in definition 3.

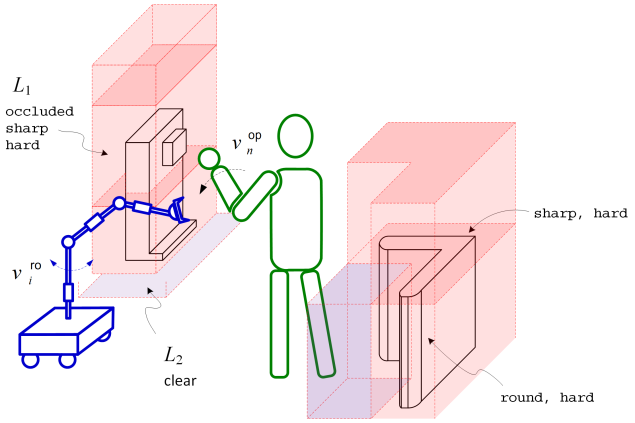


Figure 3: Example HRC layout modeled with a manipulator, an operator and some fixtures. Some locations L (shaded regions) are represented together with the values of $obst$, $shape$ and $material$ attributes.

A. Elements of HRC systems: operators, robots and layouts

The formal model of a HRC system involves three main classes: operators ($\mathcal{O}$), robots ($\mathcal{R}$) and layouts ($\mathcal{L}$).

Definition 5 (Layout ($\mathcal{L}$)): The workspace **layout** is partitioned in locations $\mathcal{L} = \{L_1, \ldots, L_n\}$, each of which is specified by a tuple of characteristics that help identify hazards and estimate their risks.

$$\begin{aligned} L_k = \langle & shape_k, material_k, obst_k \rangle \\ & shape_k \in \{\texttt{sharp}, \texttt{chamfer}, \texttt{fillet}, \texttt{round}\} \\ & material_k \in \{\texttt{soft}, \texttt{hard}\} \\ & obst_k \in \{\texttt{occluded}, \texttt{clear}, \texttt{free}, \texttt{warning}\} \end{aligned} \tag{2}$$

$shape_k$ and $material_k$ are abstract representations of the form and substance of each location and take one of the mentioned discrete values. The attributes of a location L are intended to render some safety-related properties of the environment, in order to underpin the identification of potential hazards, and some information useful for the risk estimation of such hazards (see subsection VI-A for details).

In particular, attribute $obst$ is closely related to the workspace of the robot element. Depending on where the robot is based, for each location L that the robot cannot reach (see fig. 3) it holds that $obst = \texttt{free}$. Assignments $obst = \{\texttt{occluded}, \texttt{clear}\}$ are, instead, defined for all locations within the reach of the robot, depending on the presence or absence of constraining objects. Assignment $obst = \texttt{warning}$ is defined for regions outside the robot workspace which are relevant for safety measures (e.g., slow down the robot).

For example, the layout in fig. 3 reports 3 regions around the left (hazardous) fixture, of which location L_1 is labeled as $L_1 = \langle \texttt{sharp}, \texttt{hard}, \texttt{occluded} \rangle$, while L_2 location is further away from the fixture and is labeled as $L_2 = \langle \texttt{sharp}, \texttt{hard}, \texttt{clear} \rangle$.

Definition 6 (Robot ($\mathcal{R}$)): The **robot** model $\mathcal{R}$ is composed of the formalization of its n links—that is, $\mathcal{R} = \{R_1, \ldots, R_n\}$—and several kinematic constraints that capture the correct movements of serial manipulators. A robot link R_j (e.g., an end-effector) is specified by a tuple consisting

of both static—i.e., constant over time—attributes (mass m_j and shape $shape_j$) and dynamic—whose values change over time—attributes (current values of position on the layout p_j, velocity v_j and force f_j):

$$\begin{aligned} R_j &= \langle p_j, v_j, f_j, m_j, shape_j \rangle \\ & p_j \in \mathcal{L}, \quad v_j, f_j \in \{\texttt{none}, \texttt{low}, \texttt{mid}, \texttt{high}\} \\ & m_j = \{\texttt{low}, \texttt{mid}, \texttt{high}\} \\ & shape_j \in \{\texttt{sharp}, \texttt{chamfer}, \texttt{fillet}, \texttt{round}\} \end{aligned} \tag{3}$$

As shown above, the values of velocity when translating from one location to another, and of the force exchanged between the link and the environment are expressed through quantized discrete values. For example, assuming that R_1 and R_2 are two attached links, $\mathsf{Alw}\,(p_1 = p_2 \vee \mathsf{Adj}(p_1, p_2))$ states that they are always in the same location, or in adjacent ones. Similarly, there are constraints on velocity and force values to exclude discontinuities. For example, $\mathsf{Alw}\,(v_j = \texttt{none} \Rightarrow \mathsf{Futr}\,(v_j = (\texttt{none}|\texttt{low}), 1))$ asserts that if the velocity is equal to `none`, it cannot jump to `mid` or `high` at the next time instant and it can only be equal to `low` or `none`. Another example is $\neg\mathsf{moving}_{\mathcal{R}} \Leftrightarrow \forall j(v_j = \texttt{none})$, which states that a robot is not moving when all its links have null velocity.

Remark 1: Definition 6 assumes that the position of each robot link is equal to a single location of the layout, excluding the occupation of multiple locations at the same time, and that the velocity is the center of mass velocity. Naturally, this assumption is valid only for layout partitions that are fine-grained enough. We claim this is the case for most common applications, including example 1, and the case study in section VII, as we have divided the layout into functional volumes that can be measured and are related to the programmed movements of the links. However, this assumption can be relaxed into a multi-location occupation model, which entails a greater accuracy in the modeling of the positioning of agents, but also a more complex kinematics.

Definition 7 (Operator ($\mathcal{O}$)): The model of a human **operator** is composed of a set of formalized definitions of eleven body regions $\mathcal{O} = \{O_{head}, \ldots, O_{leg}\}$ [28] and a set of constraints introduced for the sake of the consistency of the body model. Every body part O_i is defined by a tuple including both constant values which are relevant for the mechanics of contacts (mass m_i and stiffness k_i) and the dynamic values of position and velocity.

$$\begin{aligned} O_i &= \langle p_i, v_i, m_i, k_i \rangle \\ & p_i \in \mathcal{L}, \quad v_i \in \{\texttt{none}, \texttt{low}, \texttt{mid}, \texttt{high}\} \end{aligned} \tag{4}$$

For example, $\mathsf{Alw}\,(p_{head} = p_{arm} \vee \mathsf{Adj}(p_{head}, p_{arm}))$ is a constraint which asserts that consecutive body parts are always in locations that are the same or adjacent, thus ruling out unrealistic configurations such as head and arms being in opposite corners of the layout.

For the sake of compatibility with relevant standards in the field, we have chosen to adopt the human body analysis of ISO 15066 [28], which identifies eleven body regions—twenty nine specific localizations—according to their different pain tolerance on the basis of biomechanical studies in [68] on impacts (by robots). The eleven body regions are head, face, neck, shoulders, chest, belly, pelvis, upper and lower arms, hands, thigh and legs.

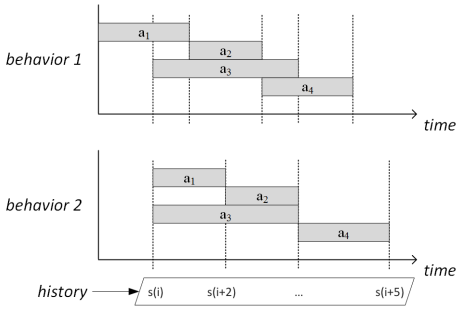


Figure 4: Sequences of actions can differ according to evaluation of their pre-conditions.

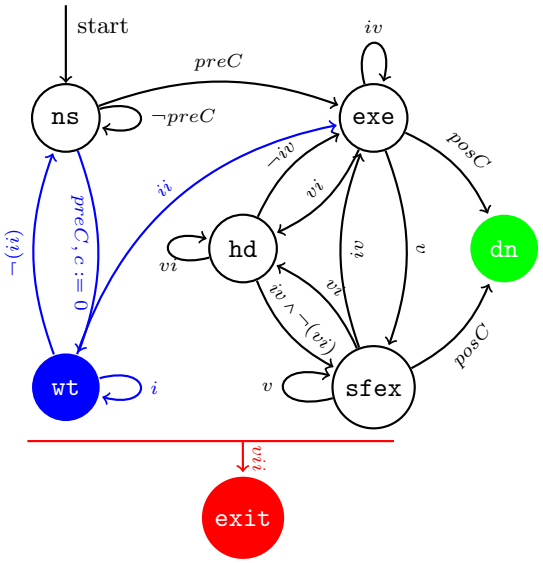
B. Formalization of tasks

In a given history H, the **state** s of the system at a given time t is the set of values taken in H by the variables, parameters and properties defined in the model: we indicate it by ${}^{H}\overline{s}(t)$. For example, $\neg\mathsf{moving}_{R_1} \wedge \mathsf{moving}_{R_2} \wedge (p_{R_1} = L_1) \wedge p_{head} = L_2$ is a constraint that predicates on the state of the model at a given time; for the constraint to hold at time t in a history H, then it must be that $v_1 = \texttt{none}$ holds at t, i.e., ${}^{H}\overline{v_1}(t) = \texttt{none}$, $v_2 \neq \texttt{none}$, etc. In this work we are interested in studying the behavior of HRC applications over *bounded histories*. More precisely, a bounded history of length K is such that only time instants $t \leqslant K$ are taken into account, with K sufficient to reach the conclusion of each history. The evolution of the model from one state to another (i.e., its dynamics) derives from *actions*, which compel attributes to change values. For example, an operator is required to move from one location (say, L_1) to another (L_2), as part of a task, while the robot is not working. The execution of this action, taking x time units, modifies the state from ${}^{H}\overline{s}(t)$, which is compatible with constraint $p_i = L_1 \wedge \neg\mathsf{moving}_{\mathcal{R}}$ to state ${}^{H}\overline{s}(t+x)$, which is instead compatible with constraint $p_i = L_2 \wedge \neg\mathsf{moving}_{\mathcal{R}}$. There might be intermediate states between instants t and $t+x$—e.g., the operator passes through other possible locations L_i and L_j to reach L_2.

Definition 8: Every task is composed of a set of atomic **actions**. An action a_i is defined by a tuple of the following form, where $a_{i,exeT}$ is its required time for termination, $a_{i,agent}$ is its expected executor agent (operator or robot), $preC_i$ and $posC_i$ are two sets of formulae that enable/control its actual execution, and $a_{i,sts}$ defines its executional state, which will be explained later on.

$$\begin{aligned} a_i &= \langle a_{i,exeT}, a_{i,agent}, preC_i, posC_i, a_{i,sts} \rangle \\ & a_{i,exeT} \in \mathbb{N}, \qquad a_{i,agent} \in \{\texttt{op}, \texttt{ro}\} \\ & a_{i,agent} = \texttt{op} \rightarrow a_{i,sts} \in \{\texttt{ns}, \texttt{wt}, \texttt{exe}, \texttt{sfex}, \texttt{hd}, \texttt{dn}, \texttt{exit}\} \\ & a_{i,agent} = \texttt{ro} \rightarrow a_{i,sts} \in \{\texttt{ns}, \texttt{exe}, \texttt{sfex}, \texttt{hd}, \texttt{dn}, \texttt{exit}\} \end{aligned} \tag{5}$$

Figure 4 shows the workflow of actions. For example, actions a_2 and a_4 could start only after termination of a_1 and a_2, respectively, hence $a_{1,sts} = \texttt{dn} \in preC_2$ and $a_{2,sts} = \texttt{dn} \in$



	Formula	Definition
i	$preC \wedge \neg$opStarts	an operator action does not start before opStarts.
ii	$preC \wedge$ opStarts $\wedge\, c \leqslant \Delta$	operator should start within Δ time units.
iii	$\exists i,j,k,x(\mathsf{RRM}_{ijk,x})$	a RRM of type x, involving R_i,O_j and L_k, is active.
iv	$\neg(iii) \wedge \neg posC$	$posC$ are not fulfilled yet and no RRM is active.
v	$(iii \wedge \neg(vi)) \wedge \neg posC$	$posC$ are not fulfilled yet and some RRM is active.
vi	$\exists i,j,k(\mathsf{RRM}_{ijk,hold})$	a RRM of type hold is active.
vii	$\exists i,j,k(\mathsf{risk}_{ijk} > \tau)$	risk exceeds the threshold.

Figure 5: Semantics of human and robot actions. For readability, some edges (numbered i-vii) are explained in the table above. For operator actions the blue part replaces the direct transition from `ns` to `exe`. c is a clock that count the time spent in state `wt`. To simplify the figure, red edges which means that risk is high and execution must abort are shown with a single edge from the set of all the states except `dn` to `exit`.

$preC_4$. However, pre-conditions do not necessarily specify a workflow as a whole: other actions may execute between a_2 and a_4, and a_2 may start on different time instants in two different histories of the model.

Model of robot actions: The semantics of our proposed model allows actions to have a set of executional states and defines the transitions among them, as depicted in fig. 5. However, this semantics does not capture the algorithm implemented by the robot controller, nor the operational state of the controller (e.g., idle, busy); rather, it is the abstraction of the conditions of robot actions. Transitions between different states of the action can also be triggered by cross-action predicates that qualify the presence of hazards, risks and RRMs. A robot action can be in one of the following states:

- `ns` (not started): initial default state which holds until all pre-conditions become true.
- `exit`: given a safety property that the system should satisfy, such as the one in definition 4, the task execution aborts upon its violation, which corresponds to the detection of hazardous situations with high risk during the lifetime of any of the actions.
- `exe` (executing): running state which is triggered when all pre-conditions are satisfied.
- `sfex` (safe executing): extension of `exe` state with at least one active RRM in order to keep the risk level acceptable.
- `hd` (hold): exception state, entered upon an explicit suspension of execution due to a request from the operator. The state is used when the execution is momentarily paused, although some safety RRMs may be enabled.
- `dn` (done): regular termination state which is triggered when all post-conditions are satisfied.

If a task contains a loop of n iterations, n separate sets of actions are independently instantiated—i.e., a_i and a_{i+n} are two separate actions with identical definitions. Note that n is a constant parameter whose value is determined before the safety analysis is carried out.

Model of operator actions: The behavior of the operator is non-deterministic. It is unrealistic to impose that the operator starts execution of an action as its preconditions are satisfied. Indeed, the execution of actions by the operator might not be as temporally accurate as for the robot. Delays could affect the correct timing of the task without compromising its completion (e.g., slight distraction, plain waiting). Hence, in addition to preconditions, an operator action depends also on a new predicate opStarts, which captures the time operator actually starts to act. As shown in formula (6), a timeout equal to Δ time units is assigned to every operator action, which starts being counted as soon as all preconditions of the action hold. The transition `wt` $\rightarrow$ `exe` is triggered when the operator starts executing the action (opStarts) within Δ time units. Otherwise, the action returns to state `ns`.

$$\begin{pmatrix} a_{i,agent} = \mathtt{op} \wedge \\ a_{i,sts} = \mathtt{wt} \end{pmatrix} \Rightarrow \begin{pmatrix} \mathsf{WithinF}\,(a_{i,sts} = \mathtt{exe}, \Delta) \vee \\ \mathsf{Futr}\,(a_{i,sts} = \mathtt{ns}, \Delta + 1) \end{pmatrix} \tag{6}$$

Another assumption included in the model states that, unlike the robot, the operator is able to execute up to two actions at a time. It is stated by the formula below, which defines that there cannot be three different operator actions executing simultaneously, thus keeping the maximum number of concurrently executing operator actions to two.

$$\neg\exists i,j,k \begin{pmatrix} i < j < k \wedge a_{z,agent} = \mathtt{op} \\ \bigwedge_{z \in \{i,j,k\}} (a_{z,sts} = \mathtt{exe} \vee a_{z,sts} = \mathtt{sfex}) \end{pmatrix} \tag{7}$$

Table II: Predicates used to formalize the definition of hazardous situations between R_j, O_i and L_k.

Predicate Formula	Definition
$\mathsf{Sep}_{ij} \in \{\texttt{close}, \texttt{mid}, \texttt{far}\}$	a discrete value to represent the distance of R_j and O_i. $\mathsf{Sep}_{ij} = \texttt{close}$ means that the two elements are in he same location ($p_i = p_j = L_k$); $\mathsf{Sep}_{ij} = \texttt{mid}$ means they are in adjacent locations; and $\mathsf{Sep}_{ij} = \texttt{far}$ means there is at least one location in between locations of i and j.
$\mathsf{InSameL}_{ijk} \Leftrightarrow \mathsf{Sep}_{ij} = \texttt{close}$	R_j and O_i are both located in L_k.
$\mathsf{ArrivedBefore}_{ijk} \Leftrightarrow \mathsf{InSameL}_{ijk} \wedge \mathsf{Past}\,(p_i = L_k, 1) \wedge \mathsf{Past}\,(p_j \neq L_k, 1)$	R_j entered L_k right before O_i.
$\mathsf{Reach}_{ijk} \Leftrightarrow \mathsf{InSameL}_{ijk} \wedge \mathsf{Past}\,(\mathsf{Sep}_{ij} > close, 1)$	R_j and O_i just arrived in L_k together.
$\mathsf{Leave}_{ijk} \Leftrightarrow \mathsf{Past}\,(\mathsf{InSameL}_{ijk}, 1) \wedge p_i \neq L_k \wedge p_j \neq L_k$	R_j and O_i both just left L_k.
$\mathsf{Contact}_{ijk} \Leftrightarrow \mathsf{InSameL}_{ijk} \wedge \mathsf{moving}_{R_i}$	contacts occur when R_j is functioning in close proximity with O_i in L_k ($\neg(L_{k,obst} = \texttt{free})$).

If multiple operator actions are in state `wt`, at most two of them will non-deterministically—and not necessarily simultaneously—enter state `exe` within Δ time units. Moreover, a human operator could manifest erroneous behavior that affects the completion of the task. This includes: repetitions of the same action, omissions, reversal of order, replacement, insertion [69]. The inclusion of state `wt` helps us to generate some of these error types (omission, insertion). This issue has been tackled more thoroughly in [70].

VI. Risk Assessment Modules

In order to automate the risk assessment procedure of ISO 12100 [1], the formal verification routine of SAFER-HRC explores all histories H of the model in search for hazards and their associated risks. Each phase of risk assessment is modeled by one of the modules explained in the following. If a history of the model follows the definition of a hazardous situation declared in subsection VI-A, then its risk is estimated by the formula included in subsection VI-B. If the estimated risk value is above a predefined threshold, the mechanisms introduced in subsection VI-C would find it and thus highlight the necessity of some RRM. Then, a suitable constraint as defined in subsection VI-D is introduced in the model.

The formal model of RRMs provides an abstraction of safety functions [71]. A safety function is deployed in practice to "solve" a non-negligible risk, and can be implemented in different ways, using various input, output and logic components (e.g., sensors, controllers, motors and mechanical devices) whose reliability level is identified by functional safety requirements. The behavior (correct or faulty) of such hardware/software components is modeled through the information they generate: for example, a correct detection by a position sensor is captured by a predicate representing the information "operator one is in region L_1" being true; Depending on the hazards detected by the verification runs, different selections of RRMs are possible, according to the chosen policy for solving risks.

If the model fails to satisfy a certain safety property, it means the model requires some refinements in terms of adding new RRMs, modifying properties of $\mathcal{O}$, $\mathcal{R}$ and $\mathcal{L}$ or pre/post-conditions of actions.

A. Hazard Identification Module

This work focuses on mechanical hazards (e.g., impacts, shearing, crushing, cutting [1, Annex B]). The principle is nevertheless extensible to other types of hazards, once they are formalized through their properties and relevant attributes. Each mechanical hazard has a formal definition and is identified by a Boolean predicate hzd^{*}_{ijk}, where $*$ is the placeholder for the hazard type label to be qualified, and ijk are, respectively, the identifiers of the robot part, human body part, and location involved in the hazardous situation. The formal model includes a finite set of hazard definitions. Whenever one of them is satisfied in a state, the corresponding predicate hzd^{*}_{ijk} is true, which captures the fact that the hazardous situation occurs in the system. Mechanical hazards arise from the relative positions of body parts, robot links and objects, as well as from the dynamics of such mutual relationships among $\mathcal{O}$, $\mathcal{R}$ and $\mathcal{L}$. Hence, some predicates have been introduced in the formal model in order to express these spatial configurations and to capture the definitions of hazards. These predicates are described in table II for given $R_j \in \mathcal{R}$, $O_i \in \mathcal{O}$ and $L_k \in \mathcal{L}$. For compatibility with recent standardization efforts in industrial robotics [28], contact hazards involving operators and collaborative robots are classified into two groups: (i) fast, impact-like contacts, where body parts are hit and then recoil because of the kinetic energy transferred to the body, defined as *transient* (`Tr`) contacts; (ii) sustained contacts of body parts against a constraining object with continuous energy flow from the robot, defined as *quasi-static* (`Qs`). In `Qs`-type situations, robot forces distribute as pressure on affected body areas with a slow/negligible displacement, which depends on the mechanical impedance of the body part. Thus, `Tr` and `Qs` hazards between R_j, O_i in L_k are formalized as follows.

Definition 9: As defined by the next formula, a **Quasi-static hazard** $\mathsf{hzd}^{\texttt{Qs}}_{ijk}$ occurs when two elements approach each other in an `occluded` location. It can also happen in a non-occlusion region when multiple links of a robot are in the same location and fold in a way that could entrap a human body part.

$$\begin{aligned} &\mathsf{hzd}^{\texttt{Qs}}_{ijk} \Leftrightarrow \\ &\mathsf{Contact}_{ijk} \wedge \mathsf{Reach}_{ijk} \wedge \mathsf{ArrivedBefore}_{ijk} \wedge \\ &\big(obst_k = \texttt{occluded} \vee \exists R_m \in \mathcal{R}(\mathsf{InSameL}_{imk} \wedge i \neq m)\big) \end{aligned} \tag{8}$$

Transient contacts are, instead, revealed by a non-sustained contact in a region without occlusions.

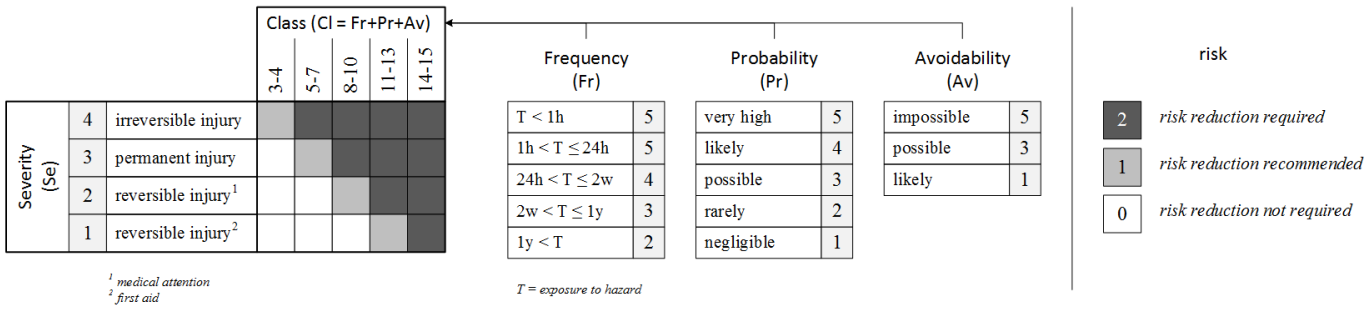


Figure 6: Reference matrix for hybrid method of risk estimation [61, §A.7, example Table A.19].

Definition 10: The next formula defines that a **Transient hazard** $\mathsf{hzd}^{\mathtt{Tr}}_{ijk}$ occurs when two elements O_i and R_j reach a `clear` location.

$$\begin{aligned}&\mathsf{hzd}^{\mathtt{Tr}}_{ijk} \Leftrightarrow\\ &\mathsf{Contact}_{ijk} \wedge \mathsf{Reach}_{ijk} \wedge \mathit{obst}_k = \mathtt{clear} \wedge\\ &\mathsf{Futr}\left(\mathsf{Leave}_{ijk}, 1\right) \wedge \nexists R_m \in \mathcal{R}(\mathsf{Reach}_{imk} \wedge j \neq m)\end{aligned} \tag{9}$$

Depending on the geometry and kinematics of the robot, the granularity with which the layout $\mathcal{L}$ is partitioned is essential to guarantee that the conditions in formulae (8) and (9) capture real hazardous situations.

Example 1 (Hazard identification): Figure 3 represents a typical HRC scenario in which the robot is supposed to execute a_n by R_j, whereas operator executes a_m by $O_i(i = hand)$. In fact the robot and operator need to **concurrently** move from L_2 to L_1 and thus $a_{m,sts} = \mathtt{exe} \in preC_n$. Recall that the layout is divided in regions $L_1 = \langle \mathtt{round}, \mathtt{hard}, \mathtt{occluded} \rangle$ and $L_2 = \langle *, *, \mathtt{clear} \rangle$, and also that $a_{n,exeT} = T_n$ and $a_{m,exeT} = T_m$ hold.
Consider a history H in which the precondition of action a_m holds at time t. The operator has Δ time units to start executing a_m and consequently different contact combinations might happen. If the timeout expires, $a_{m,sts} = \mathtt{ns}$ becomes true and a_n cannot start cause its precondition does not hold. Otherwise (the operator timely starts a_m at $t < t_1 < t + \Delta$), one of the following situations occurs:

1) At $t_2 > t_1$, O_i and R_j both reach L_2 before reaching the target (L_1). Thus, the configuration of the state at time t_2—i.e., ${}^H\overline{s}(t_2)$—matches with formula (9) and ${}^H\overline{\mathsf{hzd}^{\mathtt{Tr}}_{ij2}}(t_2)$ holds.
2) At $t_2 > t_1$, O_i arrives at L_1 before R_j. Thus, ${}^H\overline{s}(t_2)$ matches with formula (8), and ${}^H\overline{\mathsf{hzd}^{\mathtt{Qs}}_{ij1}}(t_2)$ holds.
3) O_i reaches L_1 after R_j. Thus $t_1 < t_i < t_1 + max\{T_n, T_m\}$ does not match any contact conditions.

B. Risk Estimation Module

Once the hazards are identified, their associated risks need to be computed according to the normative principles of risk estimation [1, §5.5.2]. The risk value of hzd^{*}_{ijk} is represented by risk_{ijk}, which is a compound notion of **severity** and **probability of occurrence**. Severity states the effects of hazards in terms of lesions or damages, while probability of occurrence is an indicator of the hazard's likelihood in time and space, or the possibility to being exposed to hazards due to errors and failures, or the chance to avoid hazards due to layout conditions. In case the risk is estimated to be non-negligible, RRMs should react by reducing one of the two factors—or even both, if necessary.

SAFER-HRC adopts the hybrid risk estimation method defined in ISO/TR 14121-2[61, §5.4.4.6 and Annex A] that combines a risk matrix approach with quantitative risk factors: severity Se, frequency Fr, probability Pr and avoidability Av. As shown in fig. 6, the latter three factors are combined to set an identifier ($Cl = Fr + Pr + Av$) that, together with the severity value, determines the final risk. The depicted matrix in fig. 6 replaces continuous figures (e.g., probability distributions) with qualitative notions (*likely*, *possible*, *rarely*, etc.) whose values correspond to positive integer numbers. It is noteworthy that probability scores are rarely derived from a known distribution,[2] but, rather, they are associated with statistical data about accidents in similar situations and frequent human misbehavior.

Initial risk estimation: In HRC applications, it is very common to have permanent exposure to robot systems due to the sharing of the same workspace as part of the intended use—i.e., the frequency value Fr is assumed to be always maximum and equal to the initial one Fr^0 (where the $\{\cdot\}^0$ superscript denotes the initial value of risk factors).
The probability factor Pr considers the human behavior, the reliability of components, the history of accidents, etc. The Pr score is then very much influenced by both the quality of the safety-related equipment and the skill/awareness level of the operator. In absence of any active mean intended to solve machine failures and human errors, the default score for the probability is assumed to be $Pr^0 \geqslant 2$.
Similarly, without alerting or training, nor active safety, the default chance of avoidance is negligible—i.e., $Av^0 = 5$.

Table III: Values in initial risk estimation.

	Se^0_{ijk}	Fr^0_{ijk}	Pr^0_{ijk}	Av^0_{ijk}	risk^0_{ijk}
initial value	$\geqslant 3$	5	$\geqslant 2$	5	2 (non-neg.)

The starting estimate of risks associated with hazards does not take into account any risk mitigation factor and assigns a high severity value due to the presence of moving machines

[2]This is one of the main reasons why *quantitative* methods in ISO/TR 14121-2 [61] are hardly used in manual assessments.

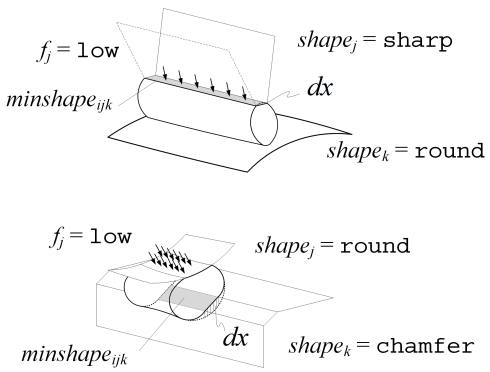


Figure 7: Pressure resulting from force f_j on surfaces $minshape_{ijk}$ (shadow area) for quasi-static contacts—O_i is always between R_j and L_k. The length of arrows is proportional to the force magnitude. The deformation of the body part under load dx is proportional to the pressure.

(default hypothesis in safety of machinery). As a result, the initial risk is very often non-negligible (see Table III), due also to its occurrence factor—i.e., for all hazard hzd^*_{ijk} it holds that $Cl^0 \geqslant 12$. Instead of adopting safeguarding RRMs, which remove hazards, but prevent the collaborative sharing of spaces and functions, HRC hazards are further analyzed in terms of their effects (i.e., the severity factor) with specific focus on physical human-robot interactions. In the following, we provide an example of more detailed model of severity of contact hazards, depending on the involved physical quantities (shape of the contact surfaces, velocity, force, etc.).

Risk Estimation: pain and pressure: The general principle in HRC physical interaction [72] is that the operator should never get injured. However, ISO 15066 claims that occasional accidental contacts would produce just a sensation of limited pain that is not different from the sensation generated by contacts in activities of daily life and could be captured by $Se < 3$. There have been several studies [73], [68], [74], [75] to set empirical injury scales. Experiments on dummies and cadavers are reported in [72]. Although the biomechanics of pain is far from being fully understood, the primary figure for painful interaction is known to be pressure, i.e., the contact force distributed through the interaction area.

Pain derives[3] from the energy or power flux density from the robot into tissues of different masses, stiffnesses and sensitivities to pain, depending on the dynamics of the contact. In particular, flux densities are dominated by kinetic energy transfer in transient contacts or by mechanical work in quasi-static constraining. The formal model contains an attribute called $minshape_{ijk} \in \{\texttt{sharp}, \texttt{chamfer}, \texttt{fillet}, \texttt{round}\}$ which is a combinatorial value of shapes of R_j, O_i and L_k and symbolizes the surface of the contact area. Once $minshape_{ijk}$ is known, the effect of the distribution of forces over surfaces and the caused pain are computable. Figure 7 shows a few examples in which the exposed area to the contact ($minshape$) affects the pressure value.

[3]from standardization perspective only. Cited ongoing studies are dedicated to investigate pain onset and bio-mechanical implications.

Table IV: The mapping between concrete and discrete f_j values, valid for human hand palm.

interaction area cm^2	$minshape$	$f_j :=$ low	$f_j :=$ mid
$5 \leqslant A < 15$	fillet	800N	3900N
$1 \leqslant A < 5$	chamfer	400N	–
$A < 1$	sharp	70N	–

When a Qs hazard occurs, the force f_j applied by the robot causes a pressure, captured by a set of discrete values $P \in \{\texttt{low}, \texttt{mid}, \texttt{high}\}$. The only case for negligible pain, formalized below, is $(P = \texttt{low})$ because it could only cause $Se < 3$. The force values in formula (10) should be initialized based on experimental pain studies. However, additional limitations could be considered if deemed necessary by the safety team. Table IV draws an example of a mapping between actual force values and the discrete values of the formal model, in the case of the palm of a human hand.

$$P^{\texttt{Qs}}_{ijk} = \texttt{low} \Leftrightarrow \tag{10}$$
$$((f_j = \texttt{low} | f_j = \texttt{mid}) \wedge (minshape_{ijk} = (\texttt{round}|\texttt{fillet})))$$
$$\vee (f_j = \texttt{low} \wedge (minshape_{ijk} = (\texttt{chamfer}|\texttt{sharp})))$$

When a Tr hazard occurs, the actual pain is believed to originate from the pressure on body tissues, whose deformation dx increases as the stiffness k_i of body part O_i decreases. Under the conservative hypothesis[4] that the kinetic energy is transformed into elastic energy, and that the momentum is preserved through the instantaneous impact, the transient contact is a totally inelastic impact, starting at relative velocity v_{ij} between R_j and O_i.

$$E = \frac{(P_i A)^2}{k_i} = \frac{m' v_{ij}^2}{2}, \qquad v_{ij} = \frac{P_i A}{\sqrt{k_i \frac{m'}{2}}} \tag{11}$$

where k_i is the stiffness of the approximate spring-damped model of the O_i.

$P^{\texttt{Tr}}_{ijk}$ captures the pain effects on the human body as a result of the compression during the impact and, like $P^{\texttt{Qs}}_{ijk}$, is discretized and formalized through a set of formulae; for example, formula (12) captures the only negligible case, which corresponds to $Se < 3$.

$$P^{\texttt{Tr}}_{ij} = \texttt{low} \Leftrightarrow (\texttt{m}' = \texttt{low} \wedge (v_{ij} = \texttt{mid}|\texttt{high})) \vee \tag{12}$$
$$((\texttt{m}' = \texttt{mid}|\texttt{high}) \wedge v_{ij} = \texttt{low}) \vee (\texttt{m}' = \texttt{low} \wedge v_{ij} = \texttt{low})$$

In conclusion, key elements for computing severity are robot force and velocity (v_j, f_j) in each state s, and the geometries of $\mathcal{R}$ and $\mathcal{L}$ rendered by $shape_j, shape_k$ attributes and ultimately expressed through predicate $minshape_{ijk}$. According to fig. 6, even the occurrences of low-severity hazards should be analyzed carefully because their combination with high probability yields non-negligible risk—i.e., risk $> \tau$. A typical example is the continuous exposure to unwanted contacts which are not tolerable for ergonomics and health reasons [76].

[4]Under the assumption that the system is isolated, any energy generated/dissipated by robot control is not modeled, so that the energy exchange between body and robot includes all the energy.

C. Risk Evaluation

A system designer should decide how conservative the system should be, and define a safety property accordingly. The global state of the model in every instant of a history should be checked against this property through an automated formal verification tool—Zot [60] in our case—to determine whether the property is violated or not. Figure 8 shows the customized output of Zot, which shows the state of the system at a time instant.

```
----------------------------time 10----------------------------
********************** Executing state **********************
Actions 1,2,3 are terminated
Action 4 is executing: operator is moving to the pallet (L_1_3)
Actions 5,6,7,8,9,10,11,12,13,14 are not started
********* Human state ***************** Robot state *********
Operator is moving          **  Robot is not moving
Head_area is in L_2_2       **  Link1 is in homing position (L_0)
Arm_area is in L_2_2        **  Link2 is in homing position (L_0)
Leg_area is in L_2_1        **  EndEff is in homing position (L_0)
*** Relative properties ********* Safety state ***
EndEff is far               **  detected hazards: none
Link1 is far                **  Risk is 0
Link2 is far                **
Relative velocity is high   **
Relative force is none      **
```

Figure 8: An example of a report on system state produced by the tool at a given time instant (t = 10).

A simplified example of a safety property is provided in formula (1). Here we introduce the actual safety property used in our approach, which captures the complexity of HRC contacts more properly. The formulae of the model explicitly state that, when the risk of a hazard reaches the acceptability threshold (τ), the associated RRM should be triggered. The purpose of defining the formula below is to check if each risk is suitably handled by at least one RRM, and it does not stay above the acceptability threshold for longer than one instant.[5]

$$\mathsf{Alw}\left(\begin{array}{l}\mathsf{risk}_{ijk} \leqslant \tau \vee \\ \mathsf{risk}_{ijk} > \tau \wedge \exists y \left(\begin{array}{l}\mathsf{RRM}_{ijk,y} \wedge \\ \mathsf{Futr}\left(\mathsf{risk}_{ijk} \leqslant \tau, 1\right)\end{array}\right)\end{array}\right) \tag{13}$$

If formula (13) is not satisfied it means that there is an RRM which is not strong enough, or that there is a hazard which has been left without mitigation. Thus, we know in order to satisfy this formula we need to define RRMs. They could be initially defined and included in the model, as we will see in subsection VI-D, or eventually come to our attention after several iterations of verification process, as it will be discussed in subsection VI-E.

D. Risk Reduction

The safety assessment approach presented in this paper supports a dynamic mechanism for activating/deactivating RRMs, so that they reduce the application's risks only when necessary. Sometimes RRMs may unacceptably downgrade the *entire* performance of the application. In general RRMs mitigate risky hazards by (i) reducing their probability (e.g., changing plan or workflow, preventing failures, etc), (ii) reducing their avoidability providing escaping conditions, (iii) reduce or eliminating the lesion on the human body in case of hazardous event. An RRM is "immediately" applied when a high level of risk is detected, and consequently, the risk value should go below the threshold "right after" the RRM starts. This is captured by the following formula where y identifies the type of the employed RRM. In our formalization "immediately" and "right after" are formalized as "simultaneously"—i.e., $\mathsf{risk}_{ijk} > \tau \wedge \exists y(\mathsf{RRM}_{ijk,y})$—and "at the next instant"—i.e., $\mathsf{Futr}\left(\mathsf{risk}_{ijk} \leqslant \tau, 1\right)$. Alternatively, one could have employed the more sophisticated notion of *zero-time transitions* [77], [78], where phenomena occur at vanishingly close time instants, which are still separate.

$$\mathsf{risk}_{ijk} > \tau \Rightarrow \exists y(\mathsf{RRM}_{ijk,y}) \tag{14}$$

The rest of this section introduces formalizations for *some* RRMs, particularly relevant for HRC situations.

Definition 11 (Speed and Separation Monitoring (SSM)): Contact-less collaborative modes can be used for protecting the operator by tuning the robot speed to maintain a minimum safe distance [29] from the operator as in the following formula. The safety team can purposefully select an RRM that tunes distance and/or velocity according to the following formula:

$$\begin{array}{l}\mathsf{RRM}_{ijk,SSM} \Rightarrow \\ (\mathsf{Sep}_{ij} < \mathsf{Sep}_{min} \Rightarrow \mathsf{Futr}\left(v_{ij} = \mathtt{none}, 1\right)) \wedge \\ (\mathsf{Sep}_{min} \leqslant \mathsf{Sep}_{ij} \leqslant \mathsf{Sep}_{mid} \Rightarrow \mathsf{Futr}\left(v_{ij} \leqslant \mathtt{mid}, 1\right))\end{array} \tag{15}$$

where Sep_{min} and Sep_{mid} are two thresholds on the distance between O_j and R_i, moving at relative speed v_{ij}. Sep_{mid} is used as a threshold for slowing down R_i, so to maintain a Sep_{min} threshold. When Sep_{min} is violated, a protection stop is triggered.

The details about the computation of relative velocity and separation distance in continuous kinematics are available in ISO/TS 15066 [28, §5.5.4]. For the consistency of formula (15) in implementations, it is essential that (i) the position of both R_j and O_i is measured by monitoring all the $\mathcal{L}$ locations, and (ii) the robot is equipped with some safety function able to limit its velocity. However, the formula is general enough to be independent from the specific implementation of the SSM strategy on actual hardware. It is able to capture either the case of keeping a constant distance through proximity sensing, considering a default human speed of 1.6m/s ISO/TS 15066[28], ISO 13855ISO:13855, or the case of dynamically updating the separation distances through online human tracking [79]. Mutual positions can be measured through any kind of environmental sensors (e.g., laser scanners) or on-board proximity sensing, while the ability to limit the velocity to a given value is part of the functions offered by the safety-related part of the robot control system (e.g., EN 61800-5-2 safely-limited speed). During the discretization of velocities for formula (15), a numerical correspondence with current technology is straightforward: robot speeds exceeding $600 - 800$ mm/s can reasonably be defined as fast for lightweight low-payload robots, whereas 2 m/s can be assumed as the fast threshold for traditional industrial medium/high-payload

[5] As a consequence of changing the property to be verified from formula (1) to formula (13), the state machine of fig. 5 needs to be adjusted in a fairly straightforward way.

robots. Similarly, $v_j = \texttt{low}$ may be assigned to $\leqslant 250$ mm/s, corresponding to a well-established value of low speed for all classes of robots.

Avoiding collisions may be hardly practicable in crowded/narrow environments, or accidental minor contacts could be unavoidable. An alternative strategy is to reduce the severity Se through, e.g., the Power and Force Limiting (PFL) collaborative mode [28, §5.5.5], derived from the recent experience in industrial domains.

Definition 12 (PFL quasi-static): The suitable RRM for a quasi-static hazard $\mathsf{hzd}^{\texttt{Qs}}_{ijk}$ is a force-limiting PFL ($\texttt{PFL}_f$ for short) protection measure:

$$\mathsf{RRM}_{ijk,\texttt{PFL}_f} \Rightarrow P^{\texttt{Qs}}_{ijk} = \texttt{low} \tag{16}$$

The effect of $\texttt{PFL}_f$ is to reduce the pressure on a body part until a safe threshold is reached, i.e., attaining the condition in formula (10). The continuous model of a constrained contact is depicted in fig. 9, with the RRM-enabling condition represented by (D). The RRM is enforced by control functions able to detect and limit the force f_i on robot links, or by intrinsically safe features (e.g., compliant actuation), or both.

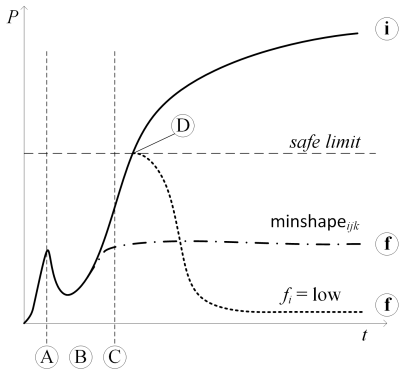


Figure 9: Typical pressure pattern (solid line) on a body part of a given surface and stiffness, in case of contact: at (A) the body loses contact because of the impact, recoils in (B), before any control action can take place after (C) (reaction time). From starting condition (i), the RRM leading to safe (f) condition can be done by limiting/detaching the robot force/torque (dotted line) or passively distributing the pressure on larger surfaces (dot-dashed line). Intrinsically safe (low-power or compliant) robots behave as the latter case.

Definition 13 (PFL dynamic): The suitable RRM for a transient hazard $\mathsf{hzd}^{\texttt{Tr}}_{ijk}$ is a speed-limiting PFL ($\texttt{PFL}_v$ for short) protection measure:

$$\mathsf{RRM}_{ijk,\texttt{PFL}_v} \Rightarrow v^{\texttt{Tr}}_{ij} = \texttt{low} \tag{17}$$

The effect of $\texttt{PFL}_v$ is to reduce the momentum transfer, and consequently the impact power through the interacting surface on body parts. Given that the reduced masses are constant, the RRM is obtained by tuning v_i, and consequently v_{ij}, as in formula (12).

E. Iterations in estimation and reduction of risks

When necessary upon risk estimation, one or more RRMs are implemented[6] by updating the properties of the system components. These updates entail that a new model is produced, corresponding to a new set of histories, hence a new formal verification activity is performed. The process explained above is done systematically in a so-called push-button way, which requires minimum manual interference to refine the model at each iteration.

Example 2: Pursuing example 1, the results of the risk analysis for identified hazardous cases can be reported as the following:

1) The RRM introduced for the hazard detected in the first case of the example (i.e., $\overset{H_1}{\overline{\mathsf{hzd}^{\texttt{Tr}}_{ijk}}}(t_2)$) is $\mathsf{RRM}_{ijk,\texttt{SSM}}$, which tackles the hazard by reducing the probability of its occurrence.
2) The hazard detected in the second case (i.e., $\overset{H_2}{\overline{\mathsf{hzd}^{\texttt{Qs}}_{ijk}}}(t_2)$) is dealt with through $\mathsf{RRM}_{ijk,\texttt{PFL}_v}$. This RRM reduces the risk by weakening its severity.

Let us emphasize that the aforementioned types of RRMs of Subsection VI-D are only some examples among several other possible ones, which have different effects on the risk levels and on the performance of the application.
The selection of RRMs in different situations depends on the safety strategy chosen by a safety analyzer or even by an automated reasoner. The result could lead to different courses of execution: one model instance may be forced to completely prevent physical contacts by implementing, e.g., collision avoidance strategies, while other options may allow physical contacts (e.g., due to the instructions of the executing action), but restraining the relative physical parameters. The incremental nature of the formal verification allows us to model both strategies (contact-tolerant vs. contact-less), to verify the formal model in either case and to enable the identification of the best strategy for various situations.

A related point to consider when choosing RRMs, in fact, is the required efficiency in terms of feasibility and execution time of a given action under the effects of a RRM. For example, an action might demand a full speed performance by the robot (e.g., no constraints are possible on velocity), or it might tolerate some level of delay caused by constraining the physical parameters of the system. The effects of the safety strategy and associated RRM selection on the verification outcome is shown trough the following example.

Example 3: Figure 10 shows an environment in which two human operators and a dual-arm robot mounted on a mobile platform work together. Assume that the robot, which is now located in location (1), should move to location (2), and then from there to (3). A collision between human (the one who is not in (1)) and robot could appear at location (X) while robot is moving from (2) to (3) and the operator is moving

[6]Functions implementing RRMs shall comply with functional safety requirements in [29, §5.5.2]. It is however out of the scope of the model and the formal verification tool to verify the compliance of safety-related parts of the control systems and all other safety functions to the functional safety standards (e.g., [71]). Each RRM is assumed to be safely deployed. See also Section VIII.

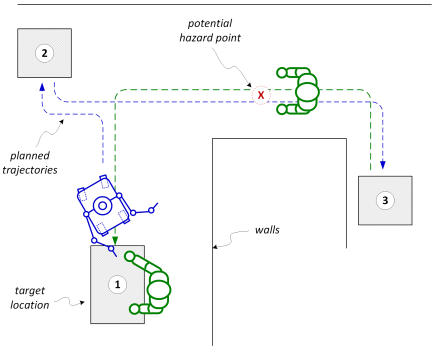


Figure 10: Example of a task involving a mobile dual-arm robot, (self)planning its motion/mission in an environment shared with human operators.

from (3) to (1). There are three possible safety strategies—and corresponding RRMs—to deal with this collision.

a) Contact-less safety strategy: If contacts are to be avoided, whenever the robot and operator's distance goes below a certain threshold, the RRM in formula (15) slows down or stops the robot. Thus, robot and operator will never meet whilst the robot is moving too fast.

b) Contact-tolerant safety strategy: Considering the layout in the figure, it seems hard to avoid contacts. Alternatively, the RRM in formula (17) could reduce the average velocity along trajectories of predictable human movements (e.g., around (X)) so that harmless physical contact occurs along colliding patterns.

c) Safety by planning/control: Another way of avoiding potentially frequent contacts, which might compromise the efficiency of the task execution, is to plan the robot to stop in (2) until the operator arrives at (1). Then it starts moving towards (3). This way the collision hazard at (X) never appears, but the execution time might considerably increase. This strategy belongs to a class of planning activities because it is a purposeful re-arrangement of the sequence of actions, redefining the intended use. The selected RRM can be modeled as a new pre-condition for action "move to (3)" of the robot. If we indicate "robot moves from (2) to (3)" as a_i, and "operator moves from (3) to (2)" as a_j then the risk reduction can be formalized as $preC_i \Rightarrow a_{j,sts} = \texttt{dn}$.

VII. Experimental Case Study

This section illustrates how SAFER-HRC works in practice by applying it to a realistic case study of collaborative fixture assembly in a Flexible Manufacturing System (FMS). The preparation of machine tool pallets—i.e., setting jigs and mounting (dismounting) workpieces into (from) fixtures before (after) machining—is a skillful and heavy job. Hence, collaborative robots are used to assist operators in tasks such as carrying/loading tools or containers, supporting workpieces during assembly, applying tools.

We consider a scenario in which the robotic system provides all services that improve the ergonomics of manual operations and release the operators from repetitive or heavy tasks. The robotic system is a manipulator arm mounted on a mobile unit that automatically relocates within a workspace as the one shown in fig. 11-(a). Regularly, the robot is positioned in front of the two assembly stations (1) and (2), or close to an instrumented inspection station (3). The robot unit can nonetheless travel and access the whole workspace, including a loading/unloading area for raw materials and finished parts, and can be manually adjusted by operators around its programmed positions (e.g., to compensate positioning/reaching limits or errors).

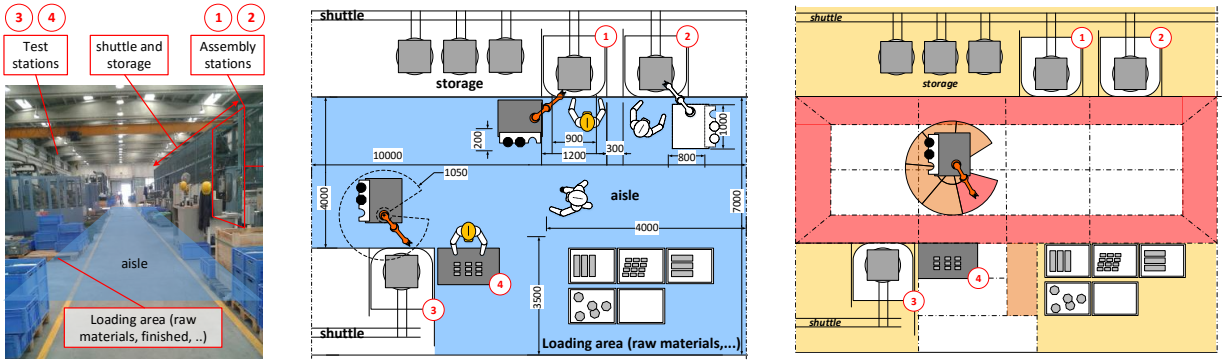
Two human operators ($\texttt{op}_1$ and $\texttt{op}_2$) are employed in the application. $\texttt{op}_1$ is mostly present on stations (1) and (2), while $\texttt{op}_2$ works mainly on (3) or executes auxiliary manual tasks on the workbench in (4). Both operators can freely hold and resume their tasks, swap posts, or join one another in some area. The main robot-assisted intended tasks are: pallet assembly at stations (1) and (2), including bin-picking from local storage carried by the mobile unit; pallet disassembly at (1) and (2), including bin-dumping; pallet inspection at station (3); lead-through programming of assembly, disassembly, and inspection tasks (trajectories, parameters, etc.) at stations (1), (2) and (3); material handling on load/unload areas.

Other manual tasks by $\texttt{op}_1$ and $\texttt{op}_2$ include manual loading of parts/boxes; (additional) visual inspection of pallet at stations (1), (2) and (3); manual assembly/disassembly of pallet at stations (1) and (2); manual measurements of parts at station (4); cleaning pallets at stations (1) and (2); kitting of tools and parts at stations (1), (2) and (3); general supervision (programming other tasks at HMI during operations, consulting production data at HMI, etc.) at stations (1), (2) and (3). Note that *all* combinations of robot/manual task assignments are admitted (e.g., robot holds and $\texttt{op}_1$ screw-drives jigs and vice versa, switching tasks on the fly, quitting a manual task and assigning the robot to proceed autonomously). Frequently, robot base and operators move side-to-side across the central aisle, or other operators transit along the aisle because the target area is part of a larger plant and access to it is not restricted. See fig. 11-(b) for reference layout and situations.

A. Model Generation

First, two instances of $\mathcal{O}$ and a $\mathcal{R}$ model with four main parts (R_1, R_2, R_{ee}, $base$) are initialized. The second step to apply our methodology is the creation of $\mathcal{L}$, which provides a discretized and abstract representation of the workspace, as in fig. 11-(c). The 3D model of the $\mathcal{L}$ workspace is rendered with stacked regions of three layers: lower, middle and upper sections. Section sizes are relative to robot and human sizes. The mobile robot base has mobility in low layers, while the manipulator arm can reach middle and upper layers, in each region. The human body is discretized in head, chest, leg, arm and fingers segments. Assuming, e.g., a person standing, the lower sections reach up to operators' hips, the middle sections cover torsos up to shoulders (where chest, arm and fingers segments are expected), and the upper sections are supposed to account for the position of operators' heads. The head can (hazardously) get down to the middle layers. Naturally, arms and fingers are the most mobile human parts.

The third step is to model the tasks, e.g., to create activity diagrams of the whole application, identify atomic actions and



(a) (b) (c)

Figure 11: (a) - The case study is carried on a real workplace, where two assembly stations and one test station are operational. The mobile robot is intended to (un)load workpieces and move across the aisle (blue area); (b) - Abstract sketch of the workspace with all the sizes and possible positions of the robot and operators; (c) - Discretized representation of the layout according to the characteristics of locations. $\mathcal{L}$ areas are colored according to their *obst* value: `occluded` (red), `clear` (orange), `free` (yellow). White areas do not have a static *obst* value, but *obst* attribute can assume `occluded`, `clear` or `free` values depending on the robot movements. The radial areas around the robotic arm actually turn around with the robot and, at any time, define the *obst* of the areas they coincide with.

formalize their pre/post-conditions. In the following, the model of tasks and their actions are described with informal language, and a sample of detailed formulae corresponding to individual actions is shown in fig. 12.

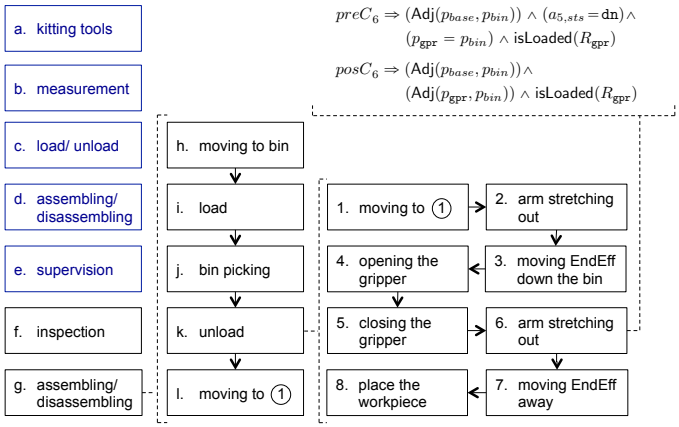


Figure 12: The case study application is composed of several tasks done by operator (a - e) and robot (f - g). `gpr` is the gripper end-effector and predicate $\text{isLoaded}(R_{\text{gpr}})$ means that the gripper is loaded with a workpiece. Here only the detail of task g is shown as an example. It is broken down into subtasks (h - l). Then, each of them in turn is a sequence of atomic actions. Here, the actions of subtask k are shown, with their intended order. The formulae shown at the top right of the figure are an example of pre/post-conditions of an action (in particular, of action 6).

The application is nominally divided in operator (i.e., a - e) and robot tasks (i.e., f - g), but different combinations of human and robot tasks can achieve the same effects. For example, executing task 'g', which is done by the robot, is equivalent to executing tasks 'a' and 'd' by the operator. Some tasks can also be switched between the operator and the robot.

B. Model Verification and Risk Reduction

As a result of a formal verification round, several samples of the SAFER-HRC output (graphical and textual traces) are shown in fig. 13. In terms hazard identification, those corresponding to time steps 17 and 24 are examples of hazards due to intended uses, while the other two show hazards caused by unintended or erroneous behavior of the operator. Time step 8 reports the occurrence of a `Tr` hazard on operator's head when he/she mistakenly bends down to execute an inspection task. Time step 20 depicts a hazardous situation resulting from a wrong manual execution (e.g., due to not following assigned instructions). For instance, the operator might confuse her/his duties and mistakenly starts to execute the task which has been assigned to the robot. Since the task is to pick pieces from the local bin, he/she is exposed to impacts.

Table V provides a numerical report of the performances of the SAFER-HRC procedure applied to the use case, which can be classified as "regular" in terms of difficulty of analysis from a safety technology and risk assessment expertise standpoint, and "challenging" in terms of extension and size of situations to consider. Unlike the general practice of risk assessment, SAFER-HRC reports *all* occurrences of hazards, including repeated/similar situations along a history: 60 instances of 27 individual hazards are found, combining the 2 base types of contact hazards (`Qs` and `Tr` impacts) along the tasks in the different required locations. In addition to being more complete than an averagely accurate risk assessment, the precise identification of hazard instances would allow a control system to trigger RRMs in due situations, at the right time.

Another benefit of the SAFER-HRC method is that a relatively large number of `Tr` hazards is identified (73% of total hazards): crushing and entanglement hazards (`Qs`) are

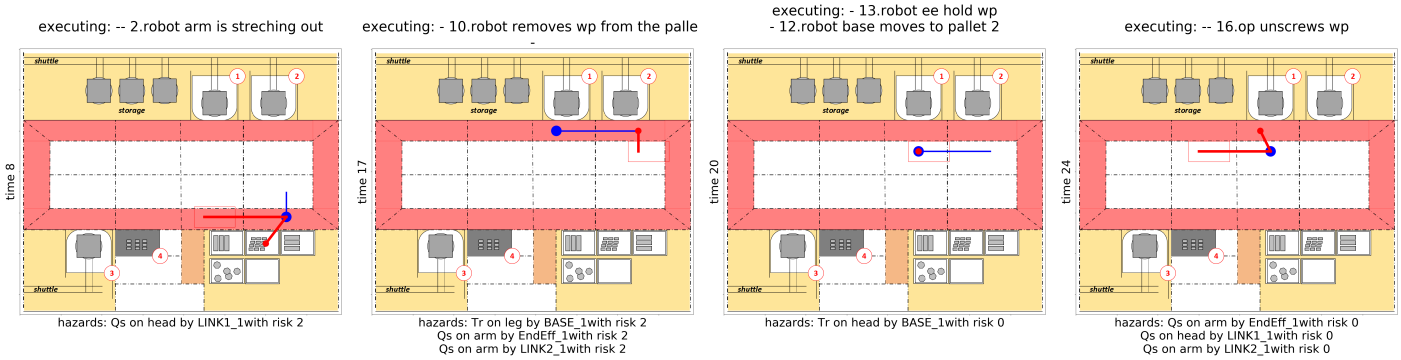


Figure 13: Examples of outputs of the tool: snapshots of different time instants.

# hazards total (individual)	# hazard types	# Qs hazards	#Tr hazards	multiple instances	# locations	# exposed body parts	support unintended use	rate of unintended use instances	# RRM types	# hazard/RRM combinations	Total required time
Hazard Identification									Risk Reduction		General
60 (27)	2	16 (27%)	44 (73%)	yes	69	5	yes	45%	5	20	20 man-hrs

Table V: Post-evaluation analysis of rounds of formal verification and model update: main performance figures in the *hazard identification* group refer to the first round (no RRMs), while figure about RRMs in *risk reduction* group refer to the application of model updates.

somehow easier to track in a HRC application because they are associated with physical locations used during task execution. Accidental dynamic impacts may instead happen all along robot movements, in any region. Such high numbers in hazard finding are due to the ability of exploring many individual locations (69) to be combined with the 5 segments used in discretizing the operators' body. Since the method encompasses also unintended behaviors, many identified hazards are derived from misuses or errors, which are an essential part of accurate risk assessments in HRC. This is highlighted by the relatively high rate of detected unintended uses (45%).

Risk reduction is achieved by manually selecting (see liability of human assessor in subsection VI-E) one of 5 possible RRMs, suitable to solve up to 20 different sub-cases of hazardous situations.

The total running time for all rounds of model verification is in the order of minutes/hours using commodity hardware,[7] to which we estimate to add several other days of model development—which might be anyway largely reused in similar applications. As a result, SAFER-HRC requires less effort than what is in general[8] necessary for completing a thorough safety assessment.

[7]The experiments were done on a Linux desktop machine with a 3.4 GHz Intel® Core™ i7-4770 CPU and 16 GB RAM of which only 235 MB have been used in the worst case. The average experiment time for each iteration was 19 minutes and 37 seconds.

[8]According to the experience of some of the authors in performing risk assessments in HRC.

VIII. Summary and Discussion

The methodology introduced in this paper incorporates, automates and improves existing standardized procedures for the safety assessment of HRC applications by using automated formal verification techniques. In particular, the methodology relies on the TRIO temporal logic for system modeling, and on the Zot formal verification tool to iteratively identify hazardous situations associated with non-negligible risks, and to mitigate them by introducing RRMs in the model.

The methodology improves manual risk assessments in terms of exhaustiveness, fallibility and accuracy, while keeping a human-in-the-loop [80] standpoint. The last feature tags this methodology as semi-automated due to the involvement of a human safety analyzer in the process of model refinement (i.e., explicitly introducing or acknowledging risk reduction models).

Discussion

This section examines in some depth the obtained results from the point of view of two relevant issues; then it addresses another important one which has been left for future research.

a) Granularity and Scalability of the model: The precision of the model is crucial for a robust identification of hazards. It might seem that the discretization of the model parameters, which is necessary for handling the size and complexity of the model, jeopardizes its precision. However, the discretization is done according to the functional specifications

and some knowledge of the task (see, e.g., the examples of levels of speed for velocity discretization in Subsection VI-D, depending on the size and payload of robots). Other notable examples of the effects of discretization concern the partitioning of layout locations, which are identified—and sized—by their application-dependent features. The discretization is designed to prevent the possibility of overlooking relevant situations (e.g., instantaneous swapping of presence in adjacent regions) and consequently missing some hazards. In Example 1 the layout region L_1 is large enough to capture the proximity of the robot and the operator—i.e., detecting the contact situations—when they move in opposite directions. This approach is generalized to any situation where the model discretization involves some critical effects [63]. However, a finer granularity increases the model complexity and verification time. Issues such as not only granularity, but also the number of layout locations, the number of elements of the robot system, the size of the task (number of actions and number of possible execution variations) can greatly increase the complexity of the model. As the model gets bigger, the formal verification tool needs more time to exhaustively explore the corresponding state-space—leading to the well known state-space explosion issue of formal verification approaches. Finding the right level of detail for models representing systems is an open issue, which inevitably involves using some level of abstraction [81]. The main consequence of increasing the complexity of the model—hence verification times, which can become hours—is the inability of executing the verification online, while the system is running. However, our experiments are based on realistic case studies, on which verification times were acceptable. Hence, we claim that the approach is effective also from the scalability point of view.

On the other hand, since this paper mainly focuses on design time the issue remains one of "trading off precision for exhaustiveness". For example, one could prefer to use a **precise simulator** that gradually generates multiple small and short traces of a system, rather than to exhaustively explore all the possible traces of a simplified **abstract system**. Both techniques have benefits and drawbacks already known to the scientific community and could be used as complementary approaches.

b) Functional Safety: The presented approach involves the use of RRMs in a purely abstract definition—i.e., as any sort of hazard-specific safety functions whose effect is to "solve" non-negligible risks (reduce the risk score). This is done for de-coupling the functional safety requirements on RRMs and the actual implementation of the safety functions that underpin RRMs. This approach has benefits in terms of formal verification, without affecting the consistency of the results obtained. For instance, the pain-pressure reduction RRM formula (10) is based on a safety function for force limitation, whose action is simply captured in the model by the condition $f_j = \texttt{low}$. Such function (i.e., formal predicate) is nonetheless implemented very differently—e.g., by tuning the maximum generated joint torque in commercial controllers, or by limiting the actuators power, or by inherently-safe design of joints (e.g., through elements of passive compliance, like SEAs). Any implementation has the effect of lowering the Se factor of the risk estimation (i.e., limit the consequences of the harm). As part of a safety function, the reliability of the risk reduction is reported by the improved probability factor, e.g., $Pr < 2$ in risk estimation. A failure in the safety function is then excluded, i.e., f_j cannot assume $\{\texttt{mid}, \texttt{high}\}$ values in the safety function predicate. When no RRM is applied, instead, failures of the system or of control functions can still be taken into account by altering the properties of the model to be verified. In the current version of the model, no requirements of functional safety are specified—e.g., no required Performance Level PL_r as for ISO 13849 is assigned and it is assumed that any RRM is implemented in the real system with the correct and suitable $PL = PL_r$. The verification of the functional safety requirements of RRMs would be nested inside the described effects of selection and application of RRMs: assigning a PL_r would be a matter of defining the exact probabilities of occurrence (derived from initial value of Pr) and gravity (derived from initial values of Se) of potential failures of the RRM. Verifying the implemented PL is a matter of computing the failure rate of the actual devices composing the safety function, together with the architecture and the Diagnostic Coverage of such function. This computation is nested because it would produce the evidence that the safety function is in fact able to improve the risk estimation factors, as captured by the formal model.

c) Stochastic Modeling: Risk estimation in HRC applications is affected by probabilistic distributions of unintended human behaviors or errors, which cumulate with distributions about failure rates or mis-functions of devices and concurrent failures/errors. In this paper we abstracted away from probabilistic analysis in the sense that we adopted some discrete values, at some level of granularity as proposed by hybrid methods [61], estimated a priori on the basis of previous experience and common sense. As pointed out in previous item VIII-0b, the Pr factor for failsafe components or functions, if deemed necessary as part of RRMs, is transformed from "high" to "low" (instead of being a numerical expression of, e.g., the ISO 13849-1 parameter PFH_d of a safety function). Associating probability distributions with detected hazard occurrences, and specifically human errors, would make the risk estimate *more accurate*, but it would not necessarily fundamentally alter the proposed approach. Indeed, probabilities could be associated with hazards after they are detected by the formal verification runs, to better estimate the actual need to introduce RRMs. In particular, we considered in [70] a series of common human errors derived from previous analyses based on psychological operator models [82]. We consider deterministic facts resulting from errors (e.g., being late, being in the wrong location, doing the wrong sequence, not respecting a post-condition). Introducing a continuous probability of such operator's errors could be the result of a suitable formalization of a parameter associated with erroneous conditions. Considering the present literature on modeling human behavior in HRC applications (e.g., the probability of repeating the same error twice, the probabilities of errors for tired, unexperienced operators with respect to expert, concentrated operators), a stochastic model of errors could help in assigning more accurate Pr factors to error appearances, but would not be formalized as a further

cause-effect relationship accounting for cognitive phenomena of any type. An objective of ongoing research is to embed complementary stochastic modeling for error parametrization into our approach based on deterministic formal verification.

ACKNOWLEDGMENT

This work was partially supported by the European Community's Seventh Framework Programme under grant agreement no. 608849 EuRoC (European Robotics Challenges) and by the Fondazione Cariplo e Regione Lombardia funding the AUTOVAM project.

REFERENCES

[1] ISO 12100, *Safety of machinery – General principles for design – Risk assessment and risk reduction*. Geneva, Switzerland: International Organization for Standardization, 2010.
[2] P. A. Lasota, T. Fong, and J. A. Shah, "A survey of methods for safe human-robot interaction," *Foundations and Trends in Robotics*, vol. 5, no. 4, pp. 261–349, 2017.
[3] IEC 60812, *Analysis techniques for system reliability – Procedure for failure mode and effects analysis (FMEA)*. Geneva, Switzerland: International Electrotechnical Commission, 2006.
[4] A. Bouti and D. A. Kadi, "A state-of-the-art review of FMEA/FMECA," *International Journal of Reliability, Quality and Safety Engineering*, vol. 01, no. 04, pp. 515–543, 1994.
[5] IEC 61025, *Fault tree analysis (FTA)*. Geneva, Switzerland: International Electrotechnical Commission, 2006.
[6] IEC 61882, *Hazard and operability studies (HAZOP studies) - Application guides*. Geneva, Switzerland: International Electrotechnical Commission, 2016.
[7] N. Leveson, "A new accident model for engineering safer systems," *Safety science*, vol. 42, no. 4, pp. 237–270, 2004.
[8] C. A. Ericson, "Fault tree analysis," in *Hazard Analysis Techniques for System Safety*. John Wiley and Sons, Inc., 2005, ch. 11, pp. 183–221.
[9] N. Leveson, *Engineering a safer world: Systems thinking applied to safety*. MIT Press, 2011.
[10] B. S. Dhillon and A. R. M. Fashandi, "Safety and reliability assessment techniques in robotics," *Robotica*, vol. 15, no. 6, pp. 701–708, 1997.
[11] G. Joshi and H. Joshi, "FMEA and alternatives v/s enhanced risk assessment mechanism," *International Journal of Computer Applications*, 2014.
[12] A. A. Baig, R. Ruzli, and A. B.Buang, "Reliability analysis using fault tree analysis: A review," *International Journal of Chemical Engineering and Applications*, 2013.
[13] J. M. Wing, "A specifier's introduction to formal methods," *Computer*, vol. 23, no. 9, pp. 8–23, Sep. 1990.
[14] J. Guiochet, Q. A. Do Hoang, M. Kaaniche, and D. Powell, "Model-based safety analysis of human-robot interactions: The MIRAS walking assistance robot," in *Proc. of the Int. Conf. on Rehabilitation Robotics (ICORR)*, 2013, pp. 1–7.
[15] D. Martin-Guillerez, J. Guiochet, D. Powell, and C. Zanon, "A UML-based method for risk analysis of human-robot interactions," in *Proc. of SERENE*. ACM, 2010, pp. 32–41.
[16] J. Guiochet, "Hazard analysis of human-robot interactions with HAZOP-UML," *Safety Science*, vol. 84, pp. 225–237, 2016.
[17] G. Booch, *The unified modeling language user guide*. Pearson Education India, 2005.
[18] N. Koenig and M. J. Matarić, "Robot life-long task learning from human demonstrations: a bayesian approach," *Autonomous Robots*, vol. 41, no. 5, pp. 1173–1188, Jun 2017.
[19] M. Webster, C. Dixon, M. Fisher, M. Salem, J. Saunders, K. L. Koay, and K. Dautenhahn, "Formal verification of an autonomous personal robotic assistant," in *Formal Verification and Modeling in Human-Machine Systems: Papers from the AAAI Spring Symposium*, 2014, pp. 74–79.
[20] M. Webster, C. Dixon, M. Fisher, M. Salem, J. Saunders, K. L. Koay, K. Dautenhahn, and J. Saez-Pons, "Toward reliable autonomous robotic assistants through formal verification: A case study," *IEEE Trans. Human-Machine Systems*, pp. 186–196, 2016.
[21] M. Sierhuis, W. J. Clancey, and R. J. V. Hoof, "Brahms: a multi-agent modelling environment for simulating work processes and practices," *International Journal of Simulation and Process Modelling*, vol. 3, no. 3, pp. 134–152, 2007.
[22] "SPIN model checker," available from spinroot.com, 2002.
[23] C. Dixon, M. Webster, J. Saunders, M. Fisher, and K. Dautenhahn, ""the fridge door is open"–temporal verification of a robotic assistant's behaviours," in *Advances in Autonomous Robotics Systems: 15th Annual Conference, (TAROS 2014). Proceedings*, 2014, pp. 97–108.
[24] R. Stocker, L. Dennis, C. Dixon, and M. Fisher, "Verifying brahms human-robot teamwork models," in *Proceedings of the European Conference on Logics in Artificial Intelligence*, 2012, pp. 385–397.
[25] J. Bredereke and A. Lankenau, "Safety-relevant mode confusions modelling and reducing them," *Reliability Engineering & System Safety*, vol. 88, no. 3, pp. 229–245, 2005.
[26] R. Butterworth, A. Blandford, and D. Duke, "Demonstrating the cognitive plausibility of interactive system specifications," *Formal Aspects of Computing*, vol. 12, no. 4, pp. 237–259, 2000.
[27] J. Fu and U. Topcu, "Synthesis of shared autonomy policies with temporal logic specifications," *IEEE Trans. Automation Science and Engineering*, vol. 13, no. 1, pp. 7–17, 2016.
[28] ISO/TS 15066, *Robots and robotic devices – Collaborative robots*. Geneva, Switzerland: International Organization for Standardization, 2016.
[29] ISO 10218-2, *Robots and robotic devices – Safety requirements for industrial robots – Part 2: Robot systems and integration*. Geneva, Switzerland: International Organization for Standardization, 2011.
[30] N. A. Stanton, "Hierarchical task analysis: Developments, applications, and extensions," *Applied Ergonomics*, vol. 37, no. 1, pp. 55–79, 2006.
[31] J. A. Marvel, J. Falco, and I. Marstio, "Characterizing task-based human-robot collaboration safety in manufacturing," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 45, no. 2, pp. 260–275, 2015.
[32] G. Jamieson, J. Andersson, A. Bisantz, A. Degani, and M. Lind, "Model-based approaches to human-automation systems design," in *Proceedings of ASME Biennial Conference on Engineering Systems Design and Analysis*, 2012, pp. 871–880.
[33] K. W. Wong, R. Ehlers, and H. Kress-Gazit, "Correct high-level robot behavior in environments with unexpected events," in *Robotics: Science and Systems X*, 2014.
[34] M. L. Bolton, E. J. Bass, and R. I. Siminiceanu, "Generating phenotypical erroneous human behavior to evaluate human-automation interaction using model checking," *Int. J. Hum.-Comput. Stud.*, vol. 70, no. 11, pp. 888–906, Nov. 2012.
[35] ——, "Using formal verification to evaluate human-automation interaction: A review," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 43, no. 3, pp. 488–503, 2013.
[36] M. L. Bolton and E. J. Bass, "Formally verifying human–automation interaction as part of a system model: limitations and tradeoffs," *Innovations in Systems and Software Engineering*, vol. 6, no. 3, pp. 219–231, 2010.
[37] J. E. Laird, *The Soar cognitive architecture*. MIT Press, 2012.
[38] J. R. Anderson, "Act: A simple theory of complex cognition." *American Psychologist*, vol. 51, 1996.
[39] D. D. Salvucci and F. J. Lee, "Simple cognitive modeling in a complex cognitive architecture," in *Proc. of CHI*, 2003, pp. 265–272.
[40] A. Cerone, P. A. Lindsay, and S. Connelly, "Formal analysis of human-computer interaction using model-checking," in *Proc. of SEFM*, 2005, pp. 352–362.
[41] P. A. Lindsay and S. Connelly, "Modelling erroneous operator behaviours for an air-traffic control task," in *Proc. of Australasian UI Conf. (AUIC)*, 2002, pp. 43–54.
[42] R. M. Young, T. R. G. Green, and T. J. Simon, "Programmable user models for predictive evaluation of interface designs," in *Proc. of CHI*, 1989, pp. 15–19.
[43] R. Butterworth, A. Blandford, and D. Duke, "The role of formal proof in modelling interactive behaviour," in *Proceedings of the Eurographics Workshop on Design, Specification and Verification of Interactive Systems (DSV-IS)*, 1998, pp. 87–101.
[44] P. Curzon, R. Rukšėnas, and A. Blandford, "An approach to formal verification of human–computer interaction," *Formal Aspects of Computing*, vol. 19, no. 4, pp. 513–550, 2007.
[45] B. Werther and E. Schnieder, "Formal cognitive resource model: Modeling of human behavior in complex work environments," in *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC)*, 2005, pp. 606–611.
[46] J. Bredereke and A. Lankenau, "A rigorous view of mode confusion," in *Proceedings of the International Conference on Computer Safety, Reliability and Security*, ser. SAFECOMP. Springer-Verlag, 2002, pp. 19–31.

[47] W. D. Gray, "The nature and processing of errors in interactive behavior," *Cognitive Science*, vol. 24, no. 2, pp. 205–248, 2000.

[48] F. E. Ritter and R. M. Young, "Embodied models as simulated users: introduction to this special issue on using cognitive models to improve interface design," *International Journal of Human-Computer Studies*, vol. 55, no. 1, pp. 1–14, 2001.

[49] C. M. Mitchell and R. A. Miller, "A discrete control model of operator function: A methodology for information display design," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 16, no. 3, pp. 343–357, 1986.

[50] H. R. Hartson, A. C. Siochi, and D. Hix, "The UAN: A user-oriented representation for direct manipulation interface designs," *ACM Trans. Inf. Syst.*, vol. 8, no. 3, pp. 181–203, Jul. 1990.

[51] F. Paternò, "Formal reasoning about dialogue properties with automatic support," *Interacting with Computers*, vol. 9, no. 2, pp. 173–196, 1997.

[52] M. L. Bolton, R. I. Siminiceanu, and E. J. Bass, "A systematic approach to model checking human-automation interaction using task analytic models." *IEEE Trans. Systems, Man, and Cybernetics, Part A*, vol. 41, no. 5, pp. 961–976, 2011.

[53] A. J. Abbate, A. L. Throckmorton, and E. J. Bass, "A formal task-analytic approach to medical device alarm troubleshooting instructions," *IEEE Trans. Human-Machine Systems*, vol. 46, no. 1, pp. 53–65, 2016.

[54] L. Feng, L. Humphrey, I. Lee, and U. Topcu, "Human-interpretable diagnostic information for robotic planning systems," in *Proc. of IROS*, 2016, pp. 1673–1680.

[55] S. Junges, N. Jansen, J. Katoen, and U. Topcu, "Probabilistic model checking for complex cognitive tasks - A case study in human-robot interaction," *CoRR*, vol. abs/1610.09409, 2016.

[56] L. Feng, C. Wiltsche, L. Humphrey, and U. Topcu, "Synthesis of human-in-the-loop control protocols for autonomous systems," *IEEE T-ASE*, vol. 13, no. 2, pp. 450–462, 2016.

[57] R. Ruksenas, J. Back, P. Curzon, and A. Blandford, "Verification-guided modelling of salience and cognitive load," *Formal Asp. Comput.*, vol. 21, no. 6, pp. 541–569, 2009.

[58] M. Askarpour, "Safer-hrc: a methodology for Safety Assessment through Formal verification in human-robot collaboration," Doctoral Dissertation, Polytechnic University of Milan, 2018, to appear.

[59] C. Baier and J.-P. Katoen, *Principles of Model Checking*. The MIT Press, 2008.

[60] "Zot: a bounded satisfiability checker," available from github.com/fm-polimi/zot, 2012.

[61] ISO/TR 14121-2, *Safety of machinery – Risk assessment – Part 2: Practical guidance and examples of methods*. Geneva, Switzerland: International Organization for Standardization, 2012.

[62] R. Pelánek, "Fighting state space explosion: Review and evaluation," in *International Workshop on Formal Methods for Industrial Critical Systems*. Springer, 2008, pp. 37–52.

[63] C. A. Furia, D. Mandrioli, A. Morzenti, and M. Rossi, *Modeling Time in Computing*, ser. Monographs in Theoretical Computer Science. An EATCS Series. Springer, 2012.

[64] R. Koymans, "Specifying real-time properties with metric temporal logic," *Real-Time Systems*, vol. 2, no. 4, pp. 255–299, 1990.

[65] M. Pradella, A. Morzenti, and P. San Pietro, "Bounded Satisfiability Checking of Metric Temporal Logic Specifications," *ACM TOSEM*, vol. 22, no. 3, pp. 20:1–20:54, 2013.

[66] E. Ciapessoni, P. Mirandola, A. Coen-Porisini, D. Mandrioli, and A. Morzenti, "From formal models to formally based methods: An industrial experience," *ACM Trans. Softw. Eng. Methodol.*, vol. 8, no. 1, pp. 79–113, 1999.

[67] L. Baresi, M. M. Pourhashem Kallehbasti, and M. Rossi, "Efficient scalable verification of LTL specifications," in *Proceedings of the 37th International Conference on Software Engineering*. IEEE Press, 2015, pp. 711–721.

[68] M. Melia, M. Schmidt, B. Geissler, J. König, U. Krahn, H. J. Ottersbach, S. Letzel, and A. Muttray, "Measuring mechanical pain: The refinement and standardization of pressure pain threshold measurements," *Behavior Research Methods*, vol. 47, no. 1, pp. 216–227, 2015.

[69] P. Curzon and A. Blandford, "From a formal user model to design rules," in *Proceedings of the International Workshop on Interactive Systems: Design, Specification, and Verification (DSV-IS), Revised Papers*, 2002, pp. 1–15.

[70] M. Askarpour, D. Mandrioli, M. Rossi, and F. Vicentini, "Modeling operator behavior in the safety analysis of collaborative robotic applications," in *Proceedings of the International Conference on Computer Safety, Reliability, and Security (SAFECOMP)*, 2017, pp. 89–104.

[71] ISO 13849-1, *Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design*. Geneva, Switzerland: International Organization for Standardization, 2006.

[72] S. Haddadin, "Physical safety in robotics," in *Formal Modeling and Verification of Cyber-Physical Systems: 1st International Summer School on Methods and Tools for the Design of Digital Systems, Bremen, Germany, September 2015*, R. Drechsler and U. Kühne, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2015, pp. 249–271.

[73] Y. Yamada, Y. Hirasawa, S. Huang, Y. Umetani, and K. Suita, "Human-robot contact in the safeguarding space," *IEEE/ASME Transactions on Mechatronics*, vol. 2, no. 4, pp. 230–236, Dec 1997.

[74] D. Mewes and F. Mauser, "Safeguarding crushing points by limitation of forces," *International Journal of Occupational Safety and Ergonomics*, vol. 9, no. 2, pp. 177–191, 2003.

[75] S. Haddadin, A. Albu-Schffer, and G. Hirzinger, "Requirements for safe robots: Measurements, analysis and new insights," *The International Journal of Robotics Research*, vol. 28, no. 11-12, pp. 1507–1527, 2009.

[76] MD 2006/42/EC, *Machinery Directive of the EC*. Geneva, Switzerland: International Organization for Standardization, 2006.

[77] E. Carpanzano, L. Ferrucci, D. Mandrioli, M. Mazzolini, A. Morzenti, and M. Rossi, "Automated formal verification for flexible manufacturing systems," *Journal of Intelligent Manufacturing*, vol. 25, no. 5, pp. 1181–1195, 2014.

[78] M. Rossi, D. Mandrioli, A. Morzenti, and L. Ferrucci, "A temporal logic for micro- and macro-step-based real-time systems: Foundations and applications," *Theoretical Computer Science*, vol. 643, pp. 38–64, 2016.

[79] F. Vicentini, M. Giussani, and L. M. Tosatti, "Trajectory-dependent safe distances in human-robot interaction," in *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, Sept 2014, pp. 1–4.

[80] P. Fung, G. Norgate, T. Dilts, A. Jones, and R. Ravindran, "Human-in-the-loop machine control loop," 1992, Patent nr. US 5116180 A.

[81] J. Guiochet, M. Machin, and H. Waeselynck, "Safety-critical advanced robots: A survey," *Robotics and Autonomous Systems*, vol. 94, pp. 43—52, 2017.

[82] E. Hollnagel, *Cognitive reliability and error analysis method (CREAM)*. Elsevier, 1998.

Federico Vicentini is a researcher of the National Research Council (CNR) at the Institute of industrial Technologies and Automation (ITIA). His research interests are in industrial robot safety, human-robot interaction and validation procedures. He serves as member of national and international standardization committees for robot and machine safety.

Mehrnoosh Askarpour is a post-doc researcher at Politecnico di Milano. Her current research interests include verification of safety-critical system properties, application of formal methods in industrial robotics and model checking for safe collaborative robotics.

Matteo Rossi is an assistant professor at Politecnico di Milano. His research interests are in formal methods for safety-critical and real-time systems, architectures for real-time distributed systems, and transportation systems both from the point of view of their design, and of their application in urban mobility scenarios.

Dino Mandrioli is a full professor of computer science at Politecnico di Milano. His research interests are in formal methods for critical systems, formal languages and their characterization in mathematical logic, automatic verification and model checking.