Hazard Analysis of Human-Robot Systems: An Adversarial Approach based on Supervisor Synthesis and Simulation

Tom P. Huck[1]

***Abstract*— Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum**

I. INTRODUCTION

Human-robot collaboration will play an important role in many areas of robotics in the future. Since close physical interactions between humans and robots can lead to hazardous situations, it is vital that the development of human-robot systems is accompanied by a thorough hazard analysis. However, analyzing hazards of HRC systems prior to commissioning is challenging. Due to close interactions and shared workflows, the human becomes a crucial safety-critical system component. While there are numerous methods to assess safety and reliability of the *technical* system components, assessing the effect of the human component is yet an unsolved problem, mainly due to the possibility of unforeseen or erroneous human behavior, which is difficult to foresee. The traditional approach when considering the impact of human behavior in hazard analyses is to make a-priori assumptions based on expert judgement. Safety standards such as ISO 12100 require that experts define a range of behaviors, including intended use and foreseeable misuse of systems, before conducting a hazard analysis. This approach is generally accepted today. Yet, as the complexity of collaborative tasks and systems continues to grow, it is becoming increasingly difficult for safety engineers to foresee what behaviors may lead to hazardous situations. If critical human behaviors are not considered in the hazard analysis of the overall human-robot system, necessary safeguards may not be implemented, or faults in the safety concept may not be identified. This leads to costly changes at later development stages and, in the worst case, accidents causing human harm.

This paper presents a model-based approach to the hazard analysis of human-robot systems. Rather than making a-priori assumptions about which behaviors are to be expected from the human, this approach treats the human as an adversarial agent who actively attempts to transfer the system into unsafe states. In doing so, the agent provides valuable examples showing how the system's safety measures can fail. This reduces the possibility of critical behaviors being overlooked.

Our approach breaks down the human-robot system into two subsystems representing human and robot, respectively. Both subsystems are modeled on two abstraction levels: Automata models capture the subsystem's interaction on an abstract level, while 3D simulation models capture the physical aspects of the interaction (e.g. sensor detection or collisions). Hazardous behaviors are synthesized automatically from the automata models on the basis of supervisory control theory. The synthesized behaviors are then simulated in a 3D simulation environment to evaluate physical safety aspects which are not sufficiently captured in the automata models.

II. PRELIMINARIES

A. Safety, Hazards and Hazard Analysis

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

B. Extended Finite State Automata (EFSA)

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet,

*This work was supported the German Federal Ministry for Economic Affairs and Climate Action under the project "SDM4FZI" (www.sdm4fzi.de). The support is gratefully acknowledged.

[1] Intelligent Process Automation and Robotics Lab, Institute of Anthropomatics and Robotics (IAR-IPR), Karlsruhe Institute of Technology, Germany. tom.huck@kit.edu

consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

C. Supervisory Control Theory

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

III. RELATED WORK

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

IV. OBJECTIVES, ASSUMPTIONS, AND PROBLEM DEFINITION

As discussed in section II, hazard analysis methods vary in terms of their methodology and the types of hazards they aim to identify. Hazards arising from random failures are *not* considered here. The justification for this omission is that there are already well-established methods to deal with random failures, and state-of-the art safety rated technology achieves extremely low failure rates (e.g., the safety integrity level SIL2, which is commonly used in safety critical robot applications, targets 10^{-7} to 10^{-6} dangerous failures per hour [1]). Instead, this paper focuses on systematic hazards, that is, hazards which arise in the interaction between human and robot due to flaws in the design of the system or its safeguards. Our approach to identify such hazards is based on the following assumptions:

(1) The system is described by two interacting subsystem models $\mathcal{H}$ and $\mathcal{R}$, with the former representing the human and the latter representing all non-human system components (i.e., robot, sensors, etc.). For brevity, the latter shall simply be called ”robot system”. The joint human-robot system shall be denoted by $\mathcal{H}||\mathcal{R}$. Concrete modeling formalisms will be introduced in Section V. For now, it is sufficient to assume that each model has a state space and and initial state.

(2) The behavior of $\mathcal{H}$ described by sequences of discrete actions from an action space $A_{\mathcal{H}}$. The set of possible sequences shall be denoted by $\mathcal{L}(A_{\mathcal{H}})$. For any given human behavior, the robot system reacts in a deterministic manner, which is encoded in $\mathcal{R}$. Hence, the behavior of the human can be regarded as the sole system input on which the behavior of $\mathcal{H}||\mathcal{R}$ depends.

(3) There is a user-defined safety specification which maps each system state to a binary value indicating if the state is safe:

$$\mathcal{SP} : S_{\mathcal{H}||\mathcal{R}} \rightarrow \{true, false\} \tag{1}$$

where $S_{\mathcal{H}||\mathcal{R}}$ is the state-space of $\mathcal{H}||\mathcal{R}$. For the system to be safe, $\mathcal{SP}$ must hold $true$ at all times during the interaction. The system is said to contain a hazard if there exists at least one state in $S_{\mathcal{H}||\mathcal{R}}$ which is (i) unsafe according to $\mathcal{SP}$ and (ii) reachable during interaction. (Note that even the state space of safe systems may contain unsafe states, however, in safe system, these states should not be reachable because the system has appropriate safeguards implemented).

Although these assumptions place certain limits on the methodology, we argue that they are justified in many practical cases. A discussion about this can be found in Section VII. For now, we take the assumptions as given and focus on their implications for the hazard analysis problem.

Observe that since the behavior of the human-robot system depends on the human behavior (assumption (2)), the model $S_{\mathcal{H}||\mathcal{R}}$ can be regarded as a function that returns resulting state trajectory $\underline{s} = (s_0, s_1, s_2, ...), s_i \in S_{\mathcal{H}||\mathcal{R}}$ for any given human action sequence $\underline{a} \in \mathcal{L}(A_{\mathcal{H}})$:

$$\mathcal{H}||\mathcal{R} : \mathcal{L}(A_{\mathcal{H}}) \rightarrow S^*_{\mathcal{H}||\mathcal{R}} \tag{2}$$

If the system contains hazards, there must therefore be at least one action sequence that transfers $S_{\mathcal{H}||\mathcal{R}}$ from the initial state to an unsafe state (assumption (3)).

We can therefore formalize the hazard analysis problem as a *search problem*: Given a model of the human $\mathcal{H}$, a model of the robot system $\mathcal{R}$, and a safety specification $\mathcal{SP}$, we search for action sequences $\underline{a} \in \mathcal{L}(A_{\mathcal{H}})$, for which the resulting state trajectory $\underline{s} = (\mathcal{H}||\mathcal{R})(\underline{a})$ contains at least one state unsafe state according to $\mathcal{SP}$.

Framing the hazard analysis in this way is an *adversarial* approach, since we deliberately aim to find human behaviors that lead to unsafe states. As discussed in Section I, this is a significant difference to traditional hazard analysis methods which generally pre-define a certain range of expected human behaviors.

V. PROPOSED SOLUTION

A. Synthesis of Adversarial Behaviors

After formalizing the problem, we turn to the question of how to find and evaluate the adversarial behaviors. Supervisor synthesis, as introduced in Section II, provides a solution to this problem, since it can systematically generate desired behaviors according to some specifications. To that end, we model $\mathcal{H}$ and $\mathcal{R}$ in the form of EFSA models. Further, we create an inverted safety specification $\overline{\mathcal{SP}}$, also in the form of an EFSA. This specification is called *inverted* since *unsafe* states are modeled as marked states (remember that, since we follow an adversarial approach, the unsafe states are desired and therefore marked). The generated supervisor, again, has the form of an EFSA. By calculating the language of the supervisor EFSA, it is now straightforward to obtain all human action sequences leading to a marked (and hence unsafe) state.

B. Simulation of Adversarial Behaviors

The behaviors and the resulting state trajectories obtained from the EFSA models in the previous step already provide some insights for safety engineers as to which behaviors may be especially critical in the analyzed human-robot system. However, EFSA models are usually rather abstract and do not allow for detailed evaluation of physical aspects such as distances or collision forces. In human-robot systems, however, such aspects are crucial in determining safety. Furthermore, there is no visualization of the critical behaviors that were identified. To address this, we perform a second evaluation step using 3D-simulation of the human-robot interactions. The supervisor now acts as a controller for the human model. Thus, the human model executes the adversarial behaviors while interacting with the robot system in simulation. This allows users to assess in a more detailed manner whether the potentially hazardous interaction scenarios identified in the previous step are actually safety-critical.

In order to further improve the level of detail of the analysis, we also consider continuous parameters of the human behavior, which are not sufficiently represented in the discrete action space of the EFSA model. This may include, for instance, continuous motion parameters such as different walking speeds, motion patterns, or sizes of the human. To consider these aspects, the supervisor executes each action sequence multiple times with different, randomly sampled parameter combinations.

VI. APPLICATION EXAMPLES AND EVALUATION

A. Example Scenario

We illustrate our approach on a simplified industrial HRC task. Further examples are available online[1], but are not covered here for reasons of brevity. The scenario is shown in Fig. x. The intended functionality of the collaborative system is as follows: The human worker starts in the center area, retreives a part from the storage area and places it in front of the robot. The worker then walks back to the center area and presses a button to activate the robot. The robot performs processing on the workpiece until the worker stops the procedure by pressing another button. The worker then retrieves the part and places it back into storage. As a safeguard against human-robot collision, the area around the robot is monitored by a laser scanner (red area in Fig x.), which sends a stop signal to the robot when the human enters the area.

In order to test the hazard analysis, the system is modified to include some safety-critical flaws. First, there is a certain delay between the time the human enters the safety zone and the stop of the robot, leading to a possible collision hazard if the human approaches sufficiently fast. Second, we introduced a safety override button by which the human can deactivate the safeguard, also leading to a collision hazard (although such an override button is unlikely to be found in a real robot system, it serves well for demonstration purposes).

B. Implementation

1) Modeling and Supervisor Synthesis: EFSA models for $\mathcal{H}$, $\mathcal{R}$, and $\mathcal{SP}$ are created using the software tool *Supremica*[verweis!]. The EFSA models are shown in Fig 1. States, events, and variables are explained in Table x. While it is not possible to explain all models in detail here, there are some interesting aspects that should be pointed out. ...

[1] Put Github link here!

Event	Explanation	Guard	Action
t_1	Transition between center and storage	-	-
t_2	Transition between center and robot	-	-
u_S	Pick up part from storage	`part_status == 0`	`part_status = 1`
d_S	Put down part at storage	`part_status == 1`	`part_status = 0`
u_R	Pick up part from robot	`part_status == 2`	`part_status = 1`; `collab_space_occ=1`
d_R	Put down part at robot	`part_status == 0`	`part_status = 2`; `collab_space_occ=1`
b_1	Press start button	`part_status != 1`	
b_2	Press safety override button	`part_status != 1`	
b_3	Press stop button	`part_status != 1`	
r	Retract hands	-	-

TABLE I

CAPTION

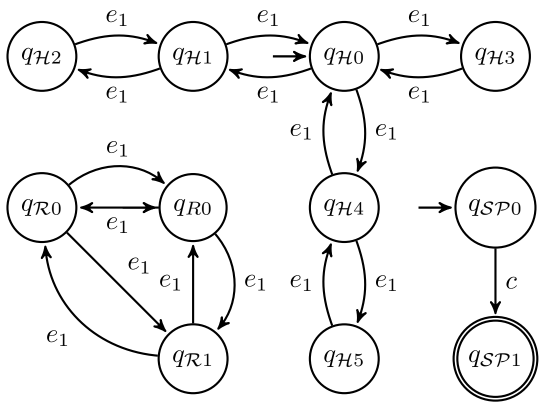

Fig. 1. EFSA models $\mathcal{H}$ (center), $\mathcal{R}$ (lower left), and safety spec $\mathcal{SP}$ (lower right).

Supervisor synthesis is also performed in supremica, which provides automated functionality for this task [verweis!].

2) Simulation: Simulations are performed in *CoppeliaSim* simulator. An image of the simulation scene is shown in Fig. x. For the human model, we use CoppeliaSim's default human model *Bill*. In order to model the inherent variability of human motion, several parameters human motion are randomized in the simulator (see Table x).

Since the safety specification $\mathcal{SP}$ only makes a binary safety decision, an additional metric is evaluated which quantifies the level of danger that is associated with a given human-robot configuration in a continuous value r.

$$r = \begin{cases} 0 & \text{case (a): } v_R < v_{crit} \\ \mathrm{e}^{-d_{HR}} & \text{case (b): } v_R \geq v_{crit};\ d_{HR} > 0 \\ \frac{F_\mathrm{c}}{F_\mathrm{max}} + 1 & \text{case (c): } v_R \geq v_{crit};\ d_{HR} = 0 \end{cases} \tag{3}$$

After the analysis, this metric allows the user to spot hazardous situations without needing to inspect each simulation run individually.

C. Comparison to other Methods

D. Results

VII. DISCUSSION

A. Reliability

- Different Lvls. of Abstraction and their implications; True/False Positives/Negatives; Overapproximate safety criteria - Modeling Effort, Modeling Bias - Synthesis vs. Search-based Testing; Degree of required knowledge (black box, white box). - Assumptions and their limitations - Level of detail, human model

B.

VIII. SUMMARY AND FUTURE WORK

ACKNOWLEDGMENT

The authors thank Tamim Asfour for his support.

REFERENCES

[1] IEC, "IEC 61508-1:2010-1 Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 1: General requirements," International Electrotechnical Commission, 2006.