

Master Thesis

Deggendorf Institute of Technology, Deggendorf

Faculty of Mechanical and Mechatronics Engineering

Master Mechatronics and Cyberphysical Systems

Framework für die sichere und zuverlässige Integration von künstlicher
Intelligenz in Leichtbauroboter

**Framework for the Safe and Reliable Integration of Artificial
Intelligence in Lightweight Robots**

Master Thesis to obtain Academic Degree

Master of Engineering(M.Eng)

submitted by: Abraham Varghese, 22101532

first examiner: Prof. Dr.-Ing, Peter Firsching

Deggendorf,08.05.2024

Confidential Disclosure Agreement

between

Deggendorf Institute of Technology

Campus Deggendorf
Dieter-Görlitz-Platz 1,
94469, Deggendorf

Faculty of Mechanical Engineering and Mechatronics
Major: Mechatronics and Cyberphysical Systems

Prof. Dr.-Ing, Peter Firsching

(in the following "Deggendorf Institute of Technology")

and

Abraham Varghese

(in the following "NEURA Robotics GmbH")

(in the following singularly and jointly "Contractual Partner")

Preamble

The Deggendorf Institute of Technology supervises an examination paper with the topic of **Framework for the Safe and Reliable Integration of Artificial Intelligence in Lightweight Robots**

(in the following "examination paper"), in which, among other things, confidential information of the company is processed. Simultaneously, confidential information also shared with the company in the context of supervision by the Deggendorf Institute of Technology.

Declaration

Name of the Student: Abraham Varghese


Name of the first Examiner: Prof. Dr.-Ing, Peter Firsching

Title of master thesis:

Framework for the Safe and Reliable Integration of Artificial Intelligence in
Lightweight Robots

I hereby declare that I have written this thesis independently. I have not submitted it for any other examination purposes. I have not used other references or material than mentioned in the bibliography and I have marked all literal analogous citations.

Deggendorf, 08.05.2024


Signature of the student:

Contents

List of Figures	iv
List of Tables	v
Acknowledgement	vi
Abstract	vii
1 Introduction	1
2 Literature Review	4
2.1 Human Robot collaboration in Industrial Environments	4
2.2 Research on ISO/IEC Standards	8
2.3 Research on Perception and AI based Grasping	12
2.4 Perception in Robotics: Risk Assessment Methods from Autonomous Driving Systems	18
2.5 Risk Assessment with FMEA in Industry 4.0	23
2.6 Conclusions from the Literature Review	24
3 Methodology	25
3.1 Robot Bin Picking Application as an Industrial Use Case for the Perception- based Grasping	25
3.1.1 System Components	26
3.1.2 Operational Sequence	27
3.1.3 Control Structure	29
3.2 Risk Assessment Methods and Tools	30
3.2.1 Failure Mode Effects Analysis	31
3.2.2 System Theoretic Process Analysis	33
3.2.3 Risk Assessment using FMEA-STPA	35
3.2.4 Decision Criteria for Treatment of the failure modes	40
3.2.5 Categorisation of Risk Priority Number	43
4 Results And Discussions	44
4.1 FMEA-STPA Results	44

4.2	Safe Practices and Safe System Requirements	50
4.2.1	Safe Practices	50
4.2.2	Safe System Requirements	51
5	Conclusion and Future Prospects	55
	References	58
6	Annex	64

List of Figures

1	Types of Human Robot Collaboration in industrial environment[3]	1
2	Present and future of Industrial Human Robot collaboration[3]	5
3	Graphical view of relationships between standards[10]	10
4	The goal of SOTIF Analysis [26]	20
5	Flow chart of the Standard Operating Procedure	28
6	Control Sequence:Inputs from sensors,processing and output	30
7	Failures Identified in Standard Operating Procedure	37

List of Tables

1 Severity Rating (S) Table 41

2 Probability of Occurrence Rating (Po) Table 42

3 Probability of Detection Rating (Pd) Table 43

4 Process FMEA table for perception-based grasping Phase 1 64

5 Process FMEA table for perception-based grasping: Phase 5.1 65

6 Process FMEA table for perception-based grasping: Phase 5.3 66

Acknowledgement

I am highly indebted to NEURA Robotics GmbH, Metzingen for their guidance and constant supervision as well as for providing necessary information and resources for the Master thesis and for the support in completing the report

I express my special gratitude to Mavarick Ho, Head of Software Strategy, and Sagar Kasrung, Functional Safety Expert, for instructing, providing information about how tasks are to be done and the flow of work, and guiding me for the thesis. Besides that, I also thank all the members for their guidance and keen support at various stages of my thesis.

I am very thankful to my Academic guide Prof. Dr.-Ing, Peter Firsching for his full support and encouragement. I owe him for his timely guidance, suggestions, and very constructive criticism which contributed immensely to the evolution of my thesis.

Abraham Varghese

Abstract

The infusion of artificial intelligence (AI) into robotics has been instrumental in automating industries, ushering in the era of the 4th Industrial Revolution. Despite the rapid advancements, the absence of specific standards guiding the operations of AI-driven robots poses inherent risks. This study examines EN ISO 10218-1,2:2011 and ISO/TS 15066, revealing a need for more explicit directives on operational dos and don'ts and system requirements for AI in robotics. To address this gap, we employ Failure Mode Effect Analysis (FMEA) as a systematic tool to analyze failure modes in AI-integrated robotics. FMEA extended with STPA provides a comprehensive approach to identify hazards related to components and subsystems. Following EN IEC 60812:2018 guidelines, we assess severity, occurrence, and detectability of the failure modes, culminating in calculating Risk Priority Numbers (RPNs) for prioritizing risks and their subsequent mitigation.

While our primary focus was on perception-based grasping, our vision extends to broader applications, AI functionalities, and varied scenarios. This iterative process aims to provide a comprehensive safety framework for AI in robotics. The study contributes to the safer integration of AI in an ever-evolving technological landscape, providing a foundation for future advancements.

1 Introduction

It was the capacity of labor that propelled humanity forward, shaping the evolution of civilization and moulding its contours. In the modern world contemporary societies emerged as hybrid environments where the physical world intertwined seamlessly with digital. Human interaction found new means when virtual communication got advanced, while artificial intelligence is standing shoulder to shoulder with natural intelligence. The tools are also being evolved at this time. The tools are no longer bounded in human hands and they are not guided solely by human will. Artificial intelligence enabled autonomy for these tools granting the machines the power to operate independently, make decisions, and chart their own course. As the world stood on the edge of this paradigm shift, the need for a deeper understanding of this technology became important.

Researchers, policymakers businesses, and society at large, all joined hands recognizing technology in all its brilliance,must remain a servant to humanity. The tool must enhance work preserving its current dual nature well spring for creation and a conduit for human fulfillment. After the pandemic COVID-19 hit the world the evolution of automation and AI was accelerated, reshaping warehouses, manufacturing plants, grocery stores, and call centers. The emergence of artificial intelligence rewrote the structure of human life transferring the way we were toiled, the way we learned and the way we structured our businesses and work. The world of robotics and automation ushered in a new era of collaboration. Man and machine are now sharing tasks, space, and even rhythm of creation.[1]

Four repeated modes of collaboration of humans and robots emerged, These modes are named co-existence, sequential collaboration, co-operation, and responsive collaboration complimenting each other's strength as shown in Figure1.



Figure 1: Types of Human Robot Collaboration in industrial environment[3]

As the wheel of progress is spinning faster robotics and AI is being advanced in tandem with Big Data, Industry 4.0, and the Internet of Things. The possibilities expanded, expanding far beyond the repetition of modes, technical configurations multiplied, each tailored to the specific demands of tasks, robots, collaboration styles, and production domains.

The integration of collaborative robot systems is a significant leap forward in industrial environments. With careful consideration of the human element, organizations can harness the full potential of this technology, redefining how humans and robots collaborate for increased productivity and safety.

The International Organization of Standardization (ISO) defines a robot as an “automatically controlled, reprogrammable multipurpose manipulator, programmable in three or more axes, which can be either fixed in place or mobile for use in industrial automation applications” .[2]

The emerged hybrid production systems aim to alleviate stress and workload blending the strengths of humans and robots. Safety became the cornerstone as robots gained the ability to sense and respond to their environment. Robots can now detect the touch of human hands, adapt to the presence of human workers and even anticipate their motions and conversations by integration of AI in robotics enhancing the safety of Human robot interaction.

Human-robot interaction is a critical aspect of collaborative robot systems. It encompasses verbal and non-verbal communication, visual information, and gestures, all of which contribute to effective collaboration. In manufacturing, where tasks are often iterative and involve physical interaction, compliance with human states and intent is paramount. While substantial progress has been made in safety within Human robot interaction, challenges remain, particularly in achieving predictability. Designing robots to be tools that humans can understand and control is crucial. This requires formal conditions for system observability and controllability, as well as an understanding of the human’s ability to control complex systems.[3]

While EN ISO 10218 part 1 and 2 were amended in 2011 there were safety measures suggested for human-robot collaboration in the collaborative cell this lead to an exponential rise in the research area of HRC. Later in ISO/TS 15066, the limiting of forces in different type of robot movement and the design of the robot cells in a safe way

were described for collaborative operations of robot with human, but AI integration and introducing cognitive abilities for a collaborative robot is not encompassed in this standards and technical specification.

The research questions we try to address through this thesis are What are the potential safety risks associated with AI-Integration in human-robot collaborative environments? How do AI-driven systems interact with other robotic functionalities, and what are the implications for overall system performance? What could be a risk analysis technique for AI-integrated robotic systems? What measures can be taken to ensure seamless integration of AI into lightweight cobots?

We try to answer these questions by studying perception and perception-based application in robotics. Perception is one technology that is integrated into the robots, which enables the robots to understand their surroundings. Perception allows the robot to identify work material, human co-workers and the dynamic environment in which the robot is. AI-based grasping is one collaborative operation of the robot which can be derived from perception. The robot can grasp an object and pass to humans or humans can work on one object and then the robot picks it up or vice versa.

A process Failure Mode Effective Analysis on AI-based grasping is done in this thesis extending it with System Theory Process Analysis. AI-based grasping algorithms are studied and possible failures of these algorithms are viewed. The Hazard and Risk assessments are based on ISO/TS 12100, keeping in mind the EN ISO 10218-1,2:2011 and the technical specification ISO/TS 15066.

To increase the reliability of the perception algorithm that induces grasping, different methods are proposed as the result of the thesis. In light of the results from this analysis of risk and hazards a safety framework is formulated for AI-based grasping, which can be taken into consideration, for a generic safety standard for the integration of artificial intelligence in robots.

2 Literature Review

In the first phase of literature survey, the research was focused on Human robot collaboration in industrial environments and the advancements made in the human robot collaboration in manufacturing environments using AI integration in robotic systems.

In the second phase of the research, the existing standards, specifications and requirements for the safe robot operations are studied. Standards like ISO 12100 and ISO/IEC 23894 the standard for risk assessment for AI integration in products are also studied. In this research we focus on formulation of standards for AI integration in industrial robots, for that the risk assessment of the failures must be done. The risk assessment was done on the failures and uncertainties of perception based grasping, since perception is one technology which enables the robot its cognitive abilities and grasping involves not only perception algorithm, but motion planing and the grasping algorithm based on the input from the perception system.

The third phase of the literature survey was on AI based grasping using robot arms and the uncertainties of the algorithm. A Failure Mode Effective Analysis was done for quantifying the risk from failures of perception based grasping. For this EN IEC 60812 and ISO 21448 were used as reference. The research based on the uncertainties and failures of the perception algorithm was helpful for brainstorming about the failure modes for the FMEA process.

2.1 Human Robot collaboration in Industrial Environments

In the chapter Human Robot Collaboration in Industrial Environments of the book *The 21st Century Industrial Robot: When Tools Become Collaborators* , the authors aim to present the existing approaches on the implementation of human robot collaboration also they try to highlight the trends for the future in which seamless integration of robots as co workers for humans. According to ISO 10218-1:2011, “collaborative workspace” is a “workspace within the safeguarded space where the robot and a human can perform tasks simultaneously during production operation”.

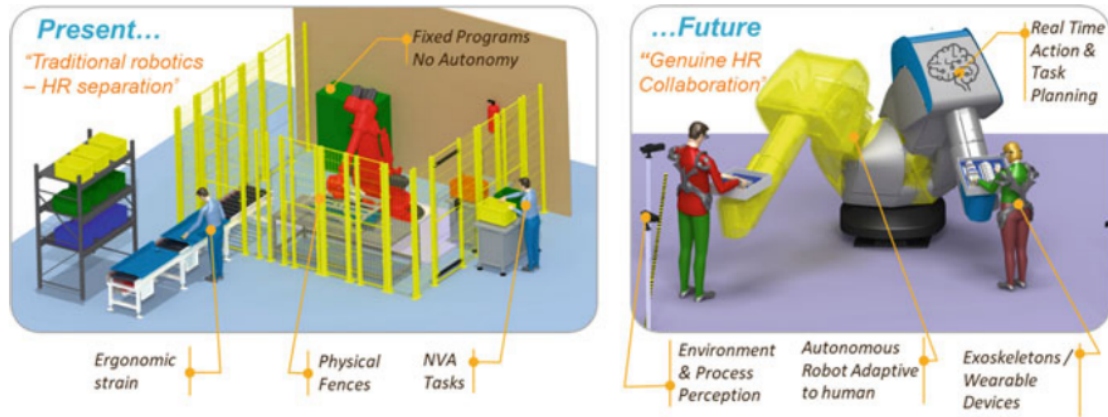


Figure 2: Present and future of Industrial Human Robot collaboration[3]

The types of human robot collaborations are explained in this chapter. The authors also explain the current state of the standards and specifications for safe industrial robots in this article. This chapter also does a Qualitative comparison between industrial robots and human operator attributes. In this article the authors are pointing out the step changes to achieve true HRC.

These are

- ”1. Development of safety related technologies that will allow industrial robots to operate in fenceless environments and/or robots that are intrinsically safe to operate around humans.
2. Deployment of appropriate robotic systems depending on the requirements of the production environments (fixed high payload robots, mobile manipulators, drones, exoskeletons).
3. Integration of Human Centred Interfaces which allow direct, natural and efficient ‘human to robot’ and ‘robot to human’ communication. The interface should be “invisible” during the interaction with the humans and ensure that physiological aspects are not negatively compromised.
4. Enablement of robotic perception to enable awareness of the tasks, environment and humans and cognition capabilities that allow them to reason upon the perceived states and adjust their behaviour for balancing task execution and interaction efficiency as a human operator would do.”[3]

The authors point out outlook on how HRC systems are expected to evolve and which are

the key enablers for an efficient and sustainable uptake of this technology. "The review has covered attempts to enable traditional robotic systems to act in a collaborative way as well as the conception and development of robots that are particularly built for the purpose. Although great leaps have been observed in individual technologies such as robotic cognition, machine vision, interfaces etc. there is still a lot of ground to be covered before seamless human robot collaboration can be achieved. From the industrial perspective and considering the technological state presented" [3]

In [4] the authors are pointing out various types of collaborative robots, manipulators and mobile, which are used in industrial environments. The authors also presents various industrial applications in which the robots are used in the industry, both in warehouse and inline production. The safety concerns are addressed in this. The introduction of sensors as precautionary methods and effectiveness of it is also discussed in this. The authors claim that the introduction of large number of the robots and the improved technologies like sensor vision, Artificial intelligence, Internet of Things enabled the 4th Industrial revolution.

In [5] The authors point out the shortcomings of the normative standards such as ISO/TS 15066 and proposes risk assessment approach resulting in a risk priority list. The author proposes FMEA and PRAT as techniques for risk assessment. A case study of robot employed for construction of brick-built residential units and help masons is done. The proposed eight steps are done on this . The eight stages proposed are Task branching, Hazard list, Body areas involved, Hazard categorization and Severity estimation, Interviews, Judgement description check, RPM computation and ranking, Risk priority list. The authors conclude after their case study that the robot cell is developed in a safe way, though the improvements for safety can be found out. This paper sets out an example for how to do risk assessment on the collaborative robot applications. Also the results of PRAT was compared with FMEA extensively used in HRC context, which are done by two different teams.

In [6] the authors posits the HRC in manufacturing application increased the task performance ,by reducing completion time and minimizing error. An extensive review of literature was done in this article how HRC is contributing to the economic motivations, occupational health and efficient use of factory space. The authors suggest about the usage of traditional industrial robots in collaborative applications , to make use of high

power of these robots. The authors discuss on the trends and future perspective of the HRC, the task of the robots in collaborative environments, the major industries in which the robots are used as the collaborative workers. Also the future trends of the market. The authors identify that since high product flexibility is possible with collaborative robots, also as the cobots are becoming cheaper in future SMEs may widely adopt cobots for wide range of industrial application.

In[7] the authors presents the advances in artificial intelligence in collaborative robots enables the robot to act in unstructured scenarios and interacting with unskilled and under trained users. Though the cobots are intrinsically safe, external sensing strategies improves safety. The research trends on increasing perception of both human and robot are addressed in this article. The standards ISO 10128 part 1 and part 2 are reviewed , the technical specification ISO/TS 15066 is also reviewed. The upcoming revision of ISO 10128 is also addressed in this review. The most relevant standards dealing with safety requirements in HRC. is listed in this article.

The ISO 691-4: 2020 provides some procedures for testing the driver-less industrial trucks designed to work automatically. This includes mobile robots in industry. The article also addresses the gaps in safety of robotic manipulators because of faster innovations and slower standardisation. The authors points out the testing and validation of safety skill required for the robot. This includes validating of safety skills like maintain safe distance, maintain dynamic stability, limit physical interaction energy, limit range of movement, maintain proper alignment, limiting restrain energy.

The authors of [8] raises their views on next generation robots , the evolution of an Human-Robot society and the categories of them, which would be there in near future are listed as industrial robots flexible for wide range of products and service robots which are capable of performing different tasks. The article raises queries about autonomy of the next generation robots and the safety issues, also about robot sociability problems. The authors point out role of AI in enabling autonomy for the robots.

The authors view about the safety standards for next generation robots are interesting in a researchers point of view for formulating the safety standards on AI integration in robotics. The article states that the difference between traditional industrial robots and next generation robots in the safety standards is that first involves machine standards and the second involves machine standards and risk from unpredictable interactions in

unstructured environments. The authors posit that the ISO 10128:2011 and the set of industrial standards addresses on the safety related parts of the control systems and the software and focuses on the robot arms and manipulator, these standards have limited application on next generation robots. Quoting the article "Complex Next Generation Robots motions, multi object interactions, and responses to shifts in environments resulting from complex interactions with humans cannot be reduced to simple performance parameters. Next Generation Robots and future Human-Based Intelligence designers and manufacturers must instead deal with unpredictable hazards associated with the legal concepts of core meaning and open texture risk." [8].

The artificial intelligence integrated robots will require a mix of pre safety and post safety mechanisms, the AI reasoning can be made for pre-safety. Safety intelligence , a system of artificial intelligence restrictions ,whose sole purpose is to provide safety when semi autonomous robot perform their tasks. So the human operator can always limit the robot autonomy. The authors suggest for clear and explicit design patterns so that semi autonomous robot can take protective reactions in human predictable ways to mitigate risks from unstable autonomous behaviour. An explicit interaction rule set and a legal architecture that can be applied to all kinds of Next Generation Robots. Also dynamic assessment of dynamic situations for response is suggested.

2.2 Research on ISO/IEC Standards

In this section of the literature research on ISO 10128:2011 part 1 and part 2, ISO/TS 15066, ISO /IEC 23894 ,EN IEC 60812 are done . When first two listed standards are for the traditional industrial robots, the 3rd standard is for the risk assessment of AI integration in different systems. IEC 60812 to provide guidelines to perform Failure Mode Effective Analysis.

Autonomous Driving is a parallel technology which can be viewed and how safety standards are implemented in the automobile software for Autonomous driving. In [9] the authors try to assess the adherence of the framework for Autonomous driving which is already in industry to ISO 26262. A case study was done on Apollo an Industrial Framework for Autonomous Driving. The observations from this article points to the need of standardisation and guidelines for GPU (Graphic processing Units) programming, which is now extensively used in processing AI algorithms. In this paper

the assessment is done and the complexities and missing features are found out and some recommendations are made for adhering the autonomous driving to ISO 26262. EN ISO 10218:2011 which is the standard for the safe use of Industrial Robots has two parts. The first part EN ISO 10218-1 :2011” specifies the requirements and guidelines for the inherent safe design protective measures and information for use of industrial robots” [2] EN ISO 10218:2011 is harmonised standard for the European Union based on EN ISO 10218 and is approved by EUROPEAN COMMITTEE FOR STANDARDIZATION (CEN). In the endorsement it is written that CEN approved the ISO 10218-1:2011 without any modification. The standard does not address the robot as a complete machine, also the standard is not applicable for non industrial robots. There are some indispensable standards for the application of this standard which are mentioned in the section Normative reference of this standard. The basic terms and definitions of these terms are explained in the clause 3 of the standard. The Hazard identification and risk assessment is the content of the 4th clause which is based on ISO 12100. There is a list of hazards that can be present in with the robots in the Annex A of the standard, further hazard identification should be done and risk assessment based on the identified hazards should be given particular consideration. Risk should be reduced or eliminated for particular scenarios like intended operation, unexpected startup, access for operators, foreseeable misuse of the robot, failure in the control system, and hazards associated with specific robot applications. The design requirements in clause 5 of the standard are based on these identified risks. These requirements are made for the safe use of the manipulator or robot arm and the design of the control system in compliance with IEC 62061:2005 or ISO 13849-1:2006. In section 10 of clause 5 there are some guidelines for the collaborative operation requirements. The stopping conditions for the robot in collaborative operations are described in this. The stop categories are in compliance with the standards IEC 60204-1. Clause 6 gives the verification and validation of the safety design and protective measures. The review of task-based risk assessment is interesting from a research point of view. 7th clause is about the information on the use of the robot. There are six annexes for the standard Annex A is already mentioned, Annex B for calculating safe distance for safeguard stop, and Annex C gives information about three position-enabling device and its functional characteristics. Annex D gives guidelines for

optional features like Stopping performance measurement, mode selection, etc. Annex E gives guidelines for labeling and Annex 6 as guidelines for the verification of safety requirements and measures.

Part 2 of ISO 10128, ISO 10128-2:2011, is made for the "recognition of the particular hazards that are presented in industrial robots when integrated and installed in industrial robot cells and lines" [10] This standard is also harmonised as EN ISO 10128-2:2011 and harmonization was without any modification by CEN.

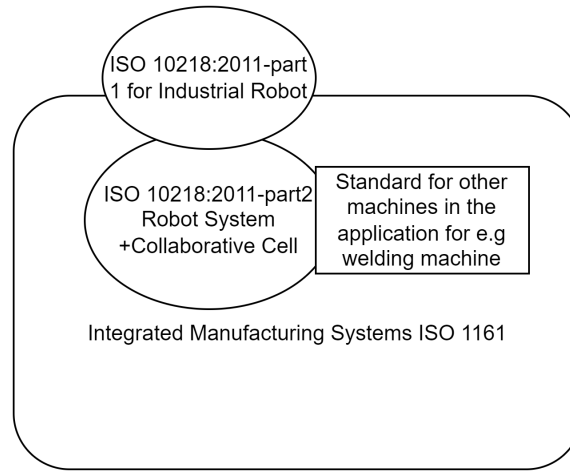


Figure 3: Graphical view of relationships between standards[10]

In the normative reference of this standard, the standards for rules and safety of different power modes and layout of the workplaces are mentioned. Clause 4 is important since it gives guidelines for Hazard identification and risk assessment when the robot is integrated into a cell. The guidelines for layout design are also mentioned in this clause. The Hazard identification guideline in this standard is based on ISO 12100 and since this is particularly for robot systems. The guidelines for task identification and assessing hazardous situations that could arise from a particular task are very useful from a research point of view. The basic design of collaborative work spaces for traditional robots is also defined in clause 5 of this standard. The Annex E of the standard describes the conceptual applications of collaborative robots, which also describes safeguards for autonomous automatic operation within a common workspace. ISO/TS 15066:2016 [11] is the technical specifications required for industrial robots for safe collaborative applications, Hazard identification, and requirements for collaborative robot operations. The conditions for the operator to enter the collaborative workspace

are specified and the power and force limiting specifications are defined. Annex A describes the Limits for quasi-static and transient contact. Maximum pressure values for quasi-static and transient contact between persons and the robot system are specified for each body part. The relationship between transferred energy and robot speed during transient contact is also mentioned in the Annex.

Even if AI is integrated into robots the movement of the robot would be based on the control systems. The control systems may take input from the outputs of the AI Algorithm. If the control system is developed according to these standards and specifications the hazards can be limited.

The ISO /IEC standard for Information Technology -Artificial Intelligence -Guidance on risk management ”gives guidance on how organizations that develop, produce, deploy or use products, systems, and services that utilize artificial intelligence (AI) can manage risk specifically related to AI.”[12] This standard in its clause 4 states the principles of AI risk management. Risk management of systems should consider the whole system. AI systems can introduce new or emergent risks. Risk management principles, frameworks, and processes to tackle this is provided in this standard. Various risk management principles applied to AI in compliance with ISO 31000 are stated in the standard. The standard gives guidance for Risk management processes. Guidelines on risk analysis are also presented in this standard. Annex B of the standard states the sources of risk, these sources must be considered when risk is analysed. These are the Complexity of the environment, Lack of transparency and explainability, Level of Automation, Risk sources related to machine learning, System hardware issues, System life cycle issues, and Technology readiness. The ISO/ IEC 23894 also refers to ISO /IEC TR 5469, which is a technical report under process for functional safety in AI systems. The International Organisation for Standardisation realizes that AI systems can introduce new safety threats thus specific standards for particular application domains should be taken into account. Annex A of the standard also mentions the seriousness of the security requirement, since data poisoning and data stealing is possible. The standard also refers to ISO / IEC 22989 which is used for establishing terminology for AI and describes concepts in the field of AI. The standard gives overall guidance for risk management of AI systems and also points to different standards that can be used for this. For, example ISO /IEC TR 24027 is a technical report on the bias and fairness of

AI systems. ISO / IEC 24029-1 for robustness of neural networks. The information from this standard is very useful for doing risk assessment for AI.

EVS-EN IEC 60812:2018[13] is a standard that explains how failure modes and effect analysis are planned, performed, documented, and maintained. "The failure mode effects and analysis is to establish how items or process might fail to perform their function so that any required treatments could be identified." [13] This standard is a generic one, this is not giving any specific applications. The basic terms and abbreviations are specified in this standard. The purposes and objectives, the roles of persons, and the skill competencies of the persons involved in the analysis are defined. The methodology is explained from planning to documentation of the process with all the guidance for the analysis. The risk priority number based on the severity occurrence and detectability of the failure is a method of assigning criticality, there are other methods mentioned for the measurement of the criticality. The standard also includes some examples of FMEA from industry applications. The standard explains the procedure and application considerations for software and processes. Annex E describes in detail the consideration for software FMEA, and the examples of failure modes are described. The other important examples given are for system-level failure causes and programming errors. The generally used columns in the spreadsheet for software FMEA are also listed in this standard. For process FMEA the starting point is a breakdown of the processes into steps. The flow diagram of the process should be analyzed. The intended outcome of each step should be defined with a sufficient description of the specific function. The standard also describes the procedure for FMEA in planning safety applications. In addition, the standard also details safety-related control systems , which can diagnose internal failures online. Before the process, FMEA was used to analyze the manufacturing process, but now it is being used in every kind of process.

2.3 Research on Perception and AI based Grasping

The motivation for doing research on perception and AI-based grasping is to study different algorithms, their unpredictability, identification of hazards, and analysis of risk. The growing significance of perception in human-robot interaction, makes robots flexible and adaptive enough for market demands.

[15] presents a comprehensive survey on robotic grasping using computer vision. Three

key tasks in grasping is object localization, object pose estimation, and grasp estimation. The explanation for object localization is given as finding the potential regions of the target objects without categorizing them. Object detection outputs the bounding boxes for the identified objects according to their categories. Object instance segmentation provides further details as pixel level or point level regions of the target. The object localization can be based on the 2D or 3D inputs and this can be fitting 2D shape primitives when the object is viewed from a fixed angle. In 2D salient object classification the arbitrary shapes are output to the detected objects. 3D localization without classification deals with 3D point cloud input, similar to 2D object localization there are 2 types described for 3D also, which are Fitting 3D shape primitives and 3D salient object detection. The shape primitives are mostly box, cylinders, or spheres. The 3D salient object detection uses RGB-D camera for depth analysis also uses fused data from the sensors for object localization. Object detection can be explained as the localization of the objects along with the classification of individual objects using a two-stage method or one-stage method. Similarly, the object instance gives the output as a detailed point cloud. Object pose estimation means finding out the 2d position and in-plane rotation angle for 2D planar grasps and 3D position and 3D orientation of the object. 6D object pose estimation is found out by transforming the object coordinate to the camera coordinate. For object pose estimation there are 3 methods described in this paper. 1) Correspondence-based methods, which is again subdivided into 2D image-based methods and 3D point cloud-based methods. In the 2D image-based method the correspondence is established through matching outputs from the RGB camera and the rendered 3D point. In 3D point cloud-based methods the correspondence is established by 3D point cloud from the RGB-D point cloud and the 3D geometric descriptors. 2) Template-based methods, involve comparison of the labeled with Ground Truth 6D object poses with obtained template. 3) Voting-based methods are further classified into Indirect voting methods and Direct voting methods. Each pixel or voxel points vote for some feature points on the correspondences. Grasp estimation means finding out the 6D pose of the gripper concerning the camera coordinate. There are 2D planar grasp and 6 DoF grasp. In a 2D planar grasp, the grasp is constrained from one direction, thus the grasp contact points can define the gripper's grasp pose. Force analytic and object geometry is important in grasp estimation, but there is only limited

data for the empirical methods of grasping causing troubles when unknown objects are being grasped. The grasp qualities are measured using Grasp quality -CNN network. In 6 DoF grasp methods are based on the point clouds and the complete shape. For 6 DoF grasp, there are different methods, methods estimating grasp qualities of candidate grasps and methods of transferring grasps from existing ones. The conclusion of the paper states the challenges, such as insufficient information in data acquisition, an insufficient amount of training data, a generalized approach in grasping novel objects, and challenges from grasping transparent objects. The adoption of multi-view data and multi-sensor data including data from haptic sensors. For tackling insufficient amounts of data the simulation environment can be built. Usage of semi-supervised learning and self-supervised learning methods to generate labeled data for 6D object pose estimation. Using plenoptic sensing.

In [16] the authors are assessing the Human-Robot perception in industrial environments. Different types of devices used for perception, especially RGB-D Camera, and the algorithms used for the functioning of these sensors along with the robot are also reviewed. The usage of these sensors in different types of robots like, fixed manipulators and mobile manipulators are also reviewed. The authors consider two scenarios of HRC with fixed manipulators. Robots have full awareness of the presence of humans. Only the awareness of the shared spaces can be sufficient, guaranteeing no human can be hurt while the robot is in motion. The usages of proprioceptive-based methods and exteroceptive methods for enhancing HRC are also reviewed. The accuracy and repetitiveness of the exteroceptive methods are still to be improved. The type of sensor, the methodology to use the sensor and the results of the methodology are the authors' interest of research. the authors made it clear that when sensor accuracy is less, the complexity of algorithms aims for compensation, to achieve an overall reliable interaction. The authors conclude based on the research that. Vision sensors are fundamental for detecting human presence in robotic systems. The data from sensor fusion can be helpful for new types of collaboration and applications. Laser sensors are also used in robotic systems in combination with vision ones mostly in the case of mobile robots. The authors also find out that, the use of non-standard sensors is still limited. In the report "Toward Safe Perception in Human-Robot Interaction" [17] the authors states perception as an important component of safety in Human-Robot Interaction. A

safe mechanical design of the robot may reduce the potential hazards, but to have a detailed knowledge of the surroundings and the state of the robot and the human operator is much beneficiary. The report suggests requirements for a holistic architecture to construct safe perception. The authors suggests the redesigning of the system is the most effective risk reduction strategy, but in less structured environment when operating adaptively redesigning alone is insufficient in most cases. To overcome this shortcoming combining of the other approaches like functional or physical safeguards and raising the awareness of the operator or user is also proposed. The report posits the use of multiple sensors and fusion of data from these sensors for redundancy. The report gives an overview about how to proceed risk analysis in safe perception. which would be very much useful from research point of view. Also suggests an architecture for safe perception for a typical collaborative robot based on the risk assessment. The usage of highly dependable sensors for perception ,at performance level D for human collaboration, for environmental perception the ToF cameras are suggested. High redundancy and heterogeneous sensors is considered as a pre-requisite for fulfilling the safety requirements.

In[18] the authors are presenting the failure modes of robotic bin picking induced from perception uncertainties. The human intervention is invoked when the robot fails . The paper describes the importance of the bin picking application using the industrial robots, since it offers a flexible automation solution. The paper explains how the complexity of bin picking increases as the shape complexity increases. Uncertainties in locating the part and the orientation of the part also leads to failure. The other perception uncertainties induced from potential of the parts for getting tangled and occlusion of grasping surfaces makes the planning challenging. Mainly the perception uncertainty leads to detection failure or singulation planning failure. The authors also give review about the related works in which uncertainties of the force, friction and contact location for grasp. The authors state they are interested in measuring the performance as a composite of "1) quality off approach of toward the object 2) grasp quality, and 3) quality of extraction of the grasped object"[18]. The paper also illustrates the confidence assessment of the robot to complete a task, so that failure of the system can be prevented, hazards and lose of money from system shutdown from these failures can also be stopped. The authors present an approach that sees

uncertainty as key to the failure of bin picking and suggests methods to deal with it. In [19] the authors try to demonstrate a framework which would take into account the overall probability of success of each grasp taking into account the error from incorrect object detection and motion error due to imperfect robot calibration. This framework takes input from multiple object detectors, grasp planners and grasp evaluators and combines these interpretations. Then uses a Bayes net model to evaluate success of each grasp. This approach can be used as a procedure for avoiding uncertainty in grasping. In [20] the authors aim to present perception challenges for grasping in clutter and unpredictable relative motion between robot and object. The authors review different types of grasping perception systems in this work. The authors investigate on performance benefits of dynamic grasping of a perception system designed to prevail the disadvantages due to occlusion, limited field of view, and minimum sensor range of a traditional wrist camera. the authors conclude stating that the placement of the perception systems is very crucial part in designing a robust robotics system. This research is focused on mobile manipulators and dynamic environment , where the manipulator and the work piece might be having high relative motion. The inputs from the research could be taken and can be upgraded, mainly about the placement of the camera, since in the company we have MAiRa Robot, which has the camera integrated in the industrial arm, and is suitable for the grasping in dynamic environment. [21] is a summary of the research challenges and Progress in Robotic grasping and Manipulation competition by IEEE Robotics and Automation Magazine. Different challenges in vision based grasping is addressed in this paper. The challenges faced in mechanism level, algorithm level and system level by the competitors are addressed. The authors tried to show the recent advancements in tackling this challenges and progress made by engineering. Also progress made by introduction of advanced hardware is also explained in this paper. The Aim of this competition is to to analyse the future research directions. The paper also gives references to benchmarks for assessment of the robustness and resilience of the mechanism. Also references for evaluation of performance of the Algorithm and Task performance by the system. The paper projects the major challenges in perception as identifying the objects with shiny surfaces, objects that are translucent and tasks which need high level of precision and accuracy. Challenges in grasping mechanisms are , grasping with imperfect perception,

since perception errors are the reason for most of the grasping failures. For example if the object's pose is not correctly identified the robot could knock over or push away the object and then improper grasping could happen. Since noise from the sensor outputs is present always the concept of perfect perception is almost impossible. Other challenges in grasping is picking up objects with complex shapes and grasping objects from clutter . Re-grasping is a major challenge since the objects needed to be adjusted after initial grasp in many of the applications. The paper also presents grasping for manipulation and in hand manipulation as very challenging areas. The authors also try to present the challenges in manipulation. The paper is concluded by pointing out, how the tasks which were considered tough, is now being solved using different approaches. Also by sharing the future scopes and ongoing researches on vision based grasp for robots. The authors in [22] tries to implement bench-marking for grasp planning algorithms. It is a very complex process since numerous factors that are diverse and complex are involved in this. The main challenges are the evaluation of grasp planning algorithm based on the influence of vision system and the arm independently. Other challenges include the difficulties confounded when doing the experiments. Lack of principle methodology that clearly defines steps in the grasping pipeline also is an issue. Standardizing this is also a major concern. The method for comparing performance of different grasp planning algorithms which can be applied for model based and model free approaches is presented in this paper. The paper presents an empirical method of verification and validation of different grasp algorithms , it explains the robot set up, the environment and the method for this verification. This method can be used to validate the algorithm , when known objects are being picked in known environments. To find out the unknown incidents that can happen in unpredictable times this method is not recommended. GRASPA 1.0 explained in [23] is a robotic grasping performance bench-marking. In this paper the environment for the benchmark is defined. Reachability within the layout camera calibration within the layout and graspability according to the maximum payloads for different robots grasp quality, execution and stability of the grasp are considered for considering the score for benchmark.

2.4 Perception in Robotics: Risk Assessment Methods from Autonomous Driving Systems

Autonomous Driving systems have more complex perception than collaborative robots since the environment for the robot is limited and more predictable than the vehicle. The stringent risk assessment methods used in AD and the safety framework can be adapted for robots. When perception-based grasping alone is considered its low-level autonomy and the sensor performance and the algorithm's performance requirement are relatively low, when medium-level autonomy, which means robots do object identification and grasping along with collision avoidance and human detection the requirements of the algorithm are relatively high.

In [24] the authors discuss the safety of perception systems in autonomous driving systems. The progress in the standardization, research advances and perspectives is also discussed. The authors are concerned about challenges in perception due to the operational conditions the edge cases and the requirements and problems of human monitoring and intervention. One of the core issues pointed out by the authors is the reliability of the black box perception systems which is outlined in the Safety of the Intended functionalities. (SOTIF). The authors also explain the main failures in perception tasks. In object detection, the two most common failures are False positives and False Negatives. False positive (FP) occurs when the object detection algorithm identifies a nonexisting object in the environment by mistake. False negatives (FN) may occur when the algorithm fails to identify an existing object in the environment. The failures occurring in object tracking can be losing track of an object due to occlusion, switching track to another object, and object re-identification. Over or under-segmentation of the scene is the most commonly occurring segmentation fault. For Autonomous Driving Systems (ADS) predictions of the object behavior are important and the failures include inaccurate predictions or missed predictions. The authors also explain the necessity of a clear framework for the development and deployment of perception systems. There are arguments stated in the paper about the perception that should address safety requirements for all ASIL attributes. ISO 21488 is further proposed for functional safety during system failures as perception systems have functional insufficiency. The standards like ANSI/UL 4600 for the safety and performance of autonomous products. The ontology for ADS taxonomy and standards

describes When it is safe for ADS and How to demonstrate system safety: based on the operating road conditions, traffic volume weather conditions, and the working state of the vehicle. The paper also explains the cause-effect chain for the intended functionality of perception, sensing, and inference which can be further used in sensor failure mitigation and sensor fusion research. Also, this helps in a challenging but very important step, quantifying the complex data-driven system's level of safety.

High-fidelity models for high-level perception systems can generate realistic results but the computational power required limits the real-time capability. the black box model to describe sensing and inference at the same time is another trend. It tries to find a direct association between final failures and the scenario. Parametrization of perception uncertainty is a goal to achieve in the future. The paper also describes the measurement of perception safety. Traditional offline safety metrics evaluate sensor inputs and outputs of ADS. Which is largely based on the ability to detect the FNs and FPs.

Uncertainty estimation is important in measuring perception safety in real-time which is based on the soft-max function as explained in the paper. This helps in providing early warnings for perception hazards. The uncertainty provides an estimate of how confident the prediction is, but they still lack any reference to real ground truth. Redundancy measures and evidence checks are used. Redundant information can be taken from different sources a pre-defined model can be used and real-time data can be cross-verified with this pre-defined model. Cooperative perception is now the trending research subject for increasing safety as per the authors. The basic idea presented is an information fusion scheme, thus increasing accuracy and reducing uncertainty. early fusion, intermediate fusion, and late fusion are implemented to achieve robust results. The authors claim that even though many cooperative perception models are being made there are new dimensions of safety concerns. This includes safety regarding communication infrastructure, local dependencies, and network-related issues. The authors also explain the V2I and V2X connection for information sharing which should be secure, for improved cooperative perception, this also can reduce edge computing nodes. The research scope of Explainable AI also sheds light on the trustable performance of the machine even when the AI model fails and is considered in the future scope of the research.

Evaluation of the safety of intended functionality is highly essential.[25] presents a

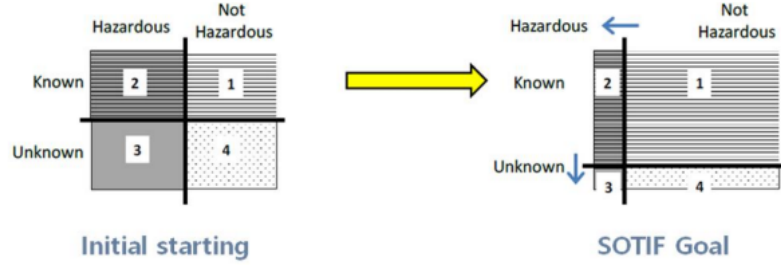


Figure 4: The goal of SOTIF Analysis [26]

concept for this. As per the authors of [25] the safety of the intended functionality is important because of the potential hazards due to the limitation of the design and implementation of the perception algorithms, and hardware performance, in SOTIF norms the limitations described are particular to the components with neural networks. In the Automotive industry, the SOTIF analysis helps in how to create or generate test cases systematically.[25] gives a contribution to evaluating the safety level of perception of the automotive system. The safety goal presented in this paper is to reduce unknown unsafe incidents and to reduce known unsafe scenarios. The authors state that risk can be calculated as the product of severity and probability, in which severity is a deterministic parameter. Probability should be found using statistical analysis. The steps like System error recovery, Partition of system failures Modelling of hazardous events, and Analysis of traffic statics are done to define test scenarios. In [26]the authors performed an analysis of how safe work can be done by an autonomous guided vehicle in a logistic environment, using SOTIF. Figure 4 shows the reduction in risks and the type of risks being reduced before and after the SOTIF analysis.

Area 3 in Figure 4 represents the unknown risks, the goal of SOTIF is to reduce Area 2 and Area 3 in the graph. These are not because of the failure but because of the impact of the surrounding environment, and limitations of situational perception according to technology performance. The authors suggest the usage of SOTIF to prevent accidents due to wrong perception or wrong performance evaluation of the algorithm in certain situations. Responsibility Sensitive Safety model based on the camera sensors is also trending in current research topics. The task of the mobile robot should be identified in a pre-requirement stage. The safety specifications should be designed according to the safety requirements identified in addition to general safety requirements. The authors

present the errors and the identified hazards in this paper for the Autonomous driving logistics robot. The paper presents a flowchart of the SOTIF process applied for the automated guided vehicle which we can try to integrate in case of robots making use of AI. The paper also presents what each clause of the SOTIF is addressing and how it was used for AGVs. The authors claim by applying the SOTIF methodology risks that can occur are eliminated as much as possible. They propose using the RSS model and SOTIF for the Design application of AGV.

In [27] the authors review and organize practical machine learning techniques that complement to the engineering safety for machine learning-based software in autonomous vehicles. the authors suggest the iterative risk assessment for the autonomous software to formally represent the operational design domain. The authors posit that the prediction probability scores in Deep Neural Networks do not provide the correct model of the uncertainty. Security challenges are there since the DNNs can be attacked and any input data alteration can fool a DNN. the paper studies ISO 26262 and ISO/PAS 21448 Standards for the Automotive vehicle. Since ISO 26262 cannot account for the faults occurring due to the the inability of the component to comprehend the environment. The authors states around 40 percentage of the software safety methods from ISO 26262 do not apply to the ML models. The Design specification is not enough for the software of the ML model because the model try to target the classes not any specification.Implementation of transparency which is specified in ISO 26262 is also not possible in ML models since ML is using high dimensional data. Testing and validation of the software is highly recommended in ISO 26262 to enforce there is no dead or unreachable code, for DNNs formally specifying their correctness is challenging. SOTIF standards treats the ML models as black box and learning problems would be there because of the error in the learning model. Run time monitoring also is an important factor that is difficult to achieve in DNN and many other machine learning models. The techniques for ML safety is stated in this paper. These techniques compliment the classic engineering strategies in software safety. To reach inherently safe AI the mankind has to travel more far. Safe fail and Safety margins must be defined. One of the practical ML safety solutions is to monitor the misclassification error detection to achieve fail safe. A parallel error detection unit should be working to monitor the transient error inputs from the hardware.Also run

time monitors should be designed three of them are listed by the authors

Uncertainty Estimation, quantifying uncertainty can explain what a model does not know in terms of model confidence on its prediction. The computation cost and hardware requirement is a challenge in this. In-Distribution Error Detectors, due to weak representation learning misclassification of in domain samples often happens, large training data sets improved DNN . Selective classification is a technique to provide high confidence samples.

Out of Distribution Error Detectors, OOD samples are the samples that are outside of the normal training distribution. OOD error is an inherent problem, revision of network architecture for learning the prediction confidence and self supervised approaches for outlier detection. Improving the algorithm's robustness is one of the the methods suggested for monitoring the machine learning algorithm. the authors propose

Robustness to domain shift and Robustness to corruptions and perturbations.

Robustness to Domain shift, Domain is the variations in the input data set compared to the training set. this reduces the performance, to overcome this domain generalization is important, unsupervised learning. Robustness to Corruptions and perturbations, data correction, and perturbation exist in the open world. Achieving model robustness to natural corruption is to improve model robustness above their clean data set.

The acceptance criteria and validation target are explained in [28] for ADS. The factors for acceptance criteria are listed as ODD factors and their attributes, Driver characteristics, Statistical factors, Relevance of Database and Values, Derivable Data, type of Rationale, possibility of multiple values, Distance vs Time, and Recentness of Data. Factors to consider for validation target are listed as Confidence level, Safety Margin, Distribution among simulation vs real-world testing, Scenario Coverage, and Multiple Acceptance Criteria, Also the authors illustrate examples in the paper for better understanding. A comprehensive results are also provided in the conclusion which we can adapt to the situations applicable to robots.

[28] suggests techniques for uncertainty evaluation in the object detection algorithms for Autonomous Vehicles. Different techniques like confusion matrix, precision-recall curve, receiver operating curve', and F-score matrix are all used in the evaluation of image classification models. The authors also suggest SOTIF analysis for the developers to test and compare the performance of the networks. The authors explain a case study for the

uncertainty evaluation.

2.5 Risk Assessment with FMEA in Industry 4.0

In [14] the authors propose FMEA-AI which is a modification for FMEA. The proposed method helps to identify safety and fairness risks in multiple failure modes of an AI system. The paper portrays "impact assessment" as an emerging mechanism to regulate AI systems. Also, the authors state that in the past AI was linked with Failure mode analysis focused on applying machine learning methods to autonomously perform FMEA, but there was no definition for identifying ways in which AI might fail. As per the author, FMEA has become a safety symbol of functional safety. The article provides two examples of doing FMEA-AI in which one is an analysis of potential applications for visual detection systems and the other one is an analysis of a series of failure modes for a single AI application. The context of severity and unfairness is explained in this article by the authors. In the proposed method people are divided into user groups to find out the unfairness. The article gives an outlook on how the probability of occurrence, risk, and mitigation is considered. The FMEA work sheet for a specific application can be used as a guideline for doing FMEA on the application AI-based perception and grasping.

In[30] the authors propose FMEA-STPA for risk analysis in intelligent and collaborative automation systems. Since this is predicted to be an important part of flexible manufacturing. This could contribute to evaluating the architectural design, and the risks and applying the risk reduction measures. While FMEA contributes to the system through reliability theory STPA contributes through system thinking. The authors explore the integration of FMEA and STPA to address the challenges that arise from the gaps in the guidelines addressing risk related to machines empowered by AI. This allows a more holistic analysis of the control structure. The promising path opened by this integration of risk assessment tools a more comprehensive evaluation of safety is ensured. In this article the authors show how the risk assessment techniques from Autonomous systems of a car complement the robotic field. The previous papers published by the authors of this paper also suggest planned safety , and its implementation [35] which is also suggested in this thesis and use-case to avoid compromising the cycle times of the application.

2.6 Conclusions from the Literature Review

The first part of the review gave ideas about the improvement of human-robot interactions using AI and how important the integration of AI in robotics is. The research on ISO/IEC Standards like EN ISO 10218-1,2:2011 gave light on the risks of robots in the industrial atmosphere. General standards like IEC 61508 and ISO 12100 gave ideas on determining the risk and methods to determine risk. The review of IEC 60812:2018 gave guidelines to do FMEA which helps to quantify risks. The ISO/TS 15066 also gives specifications for mitigating the risk of collaborative robots. The study of grasping algorithms could contribute to understanding how perception-based grasping is working and what can be the possible failures. The research on the safety of perception systems of autonomous vehicles helped to contribute to understanding the failures from the algorithms also the techniques used to mitigate them. The knowledge gained from this literature review is used as the base for the methodology of the thesis.

3 Methodology

This chapter discusses the risk assessment tool used for Artificial intelligence integration in robotics. The tool we use here is Failure Mode Effect Analysis with an extension of System Theory Process Analysis. In light of the absence of established standards within the industries and market, we use FMEA-STPA as a tool for conducting risk assessment for the AI integration in robotics, this would help to identify and mitigate the potential failures and associated risks. The identified potential failures are similar to the data of the Autonomous Driving system. We use FMEA-STPA on perception-based grasping in robots. Since it is one of the best-known scenarios where robots make use of Artificial intelligence and do manipulation tasks. The reason for selecting FMEA and adding an extension of STPA for the risk assessment and the basic definitions of the terms used in processes are explained. The system setup and the boundary conditions are explained in a subsection of the chapter. The use case from which the results are manipulated is also explained along with its relevance.

3.1 Robot Bin Picking Application as an Industrial Use Case for the Perception-based Grasping

In an ideal smart system where robots and humans work together, tasks are split into different jobs, each with smaller tasks. For example, imagine a situation where robots and people team up to pick warehouse parts, put them together a bit, and then move the assembled kits. Some tasks need both humans and robots to work together, while others need each to do their own thing. Besides the main tasks, there are other jobs like fixing problems, figuring out issues, learning new things, giving or getting instructions, handling changes, and planning what to do. Everyone in the system has to work together in these tasks, and sometimes there might be risky situations.

The bin-picking robot system is a highly advanced automation solution designed for the efficient picking and placing of machined parts from two adjacent heavy boxes onto a conveyor belt. This fixed robotic system incorporates cutting-edge technologies, including a magnetic gripper with a spring mechanism that allows robust picking, an RGB-D camera for precise object identification, and Safe Human Detection sensors for human presence detection.

3.1.1 System Components

Robot Arm: The robot arm is securely fixed on a stationary base which is a versatile and powerful solution for handling tasks with precision. With a strong payload capacity of 12-14 kg and an extensive reach of 1400 mm, it's well-equipped for various industrial applications. What makes it stand out is its intelligent head, which cleverly integrates Lidar sensors and an RGB-D camera directly into the arm's seventh axis. Unlike traditional setups, this design enhances the arm's flexibility by seamlessly incorporating advanced sensing capabilities. With seven degrees of freedom, it can move precisely in complex tasks, adapting to different orientations. Plus, it's built tough with an IP 65 rating, ensuring resilience against dust and water—perfect for challenging industrial settings. The intelligent head's Lidar sensors enable real-time 3D mapping, helping the arm navigate and plan its path effectively. The integrated RGB-D camera adds high-resolution color imagery and depth perception, making it adept at recognizing objects and handling them precisely. This robotic arm finds applications in manufacturing, assembly lines, logistics, and research and development tasks. Its adaptability, combined with a robust IP 65 rating, makes it a game-changer for organizations seeking advanced automation solutions. Equipped with a magnetic gripper as an end effector for robust and adaptable picking.

Magnetic Gripper: Features a spring mechanism for smooth and secure picking of machined parts. The magnetic force ensures a strong grip on objects, preventing unintentional drops.

RGB-D Camera: Mounted on the robot arm for comprehensive color (RGB) and depth (D) information. Utilized for the identification of both the box and the machined parts within it.

Safe Human Detection Sensors: Integrated into the system to identify the presence of humans in the operational vicinity. Provides real-time feedback for safety and operational considerations.

External PLC: The Programmable Logic Controller integrated into the system is used for the human-robot interaction addition to the graphical user interface.

Conveyor Belt: The conveyor belt is placed within reach of the robot arm, to place the object picked from the box. The conveyor belt moves the placed workpiece to the next workstation.

Control System: Manages the overall operation of the robot system and receives input from the RGB-D camera and Safe Human Detection sensors, This System controls the actions of the robot arm, magnetic gripper, and conveyor belt.

The heart of the robotic system lies within a sleek and efficient modular control box. This compact unit seamlessly combines the power of Artificial Intelligence (AI), safety protocols, motion control, and power management. The modular design streamlines the control processes, ensuring a harmonious interaction between these vital components. With AI, the robotic arm gains intelligent decision-making capabilities, adapting to various scenarios. The safety module prioritizes secure operations, and the motion control module enables precise and fluid movements. Simultaneously, the power module efficiently manages energy distribution for a reliable and continuous power supply. This modular control box represents the central intelligence that propels our cutting-edge robotic arm, ensuring optimal performance across a spectrum of applications.

3.1.2 Operational Sequence

1. **Start-Up:** Power on the entire system, including the robot arm, magnetic gripper, RGB-D camera, and Safe Human Detection sensors, Initiate the control system for seamless operation.
2. **Picking Phase:** The robot arm starts picking machined parts from the first box using the magnetic gripper.
3. **Object and Box Identification:** The RGB-D camera identifies the box and the machined parts within it with high precision, Safe Human Detection sensors continuously monitor the surroundings for the presence of humans.
4. **Orientation Check:** The system checks the orientation of the picked object. If incorrect, the robot attempts a re-pick, allowing for up to two failures before corrective actions.
5. **Lookup Point and Scene Re-identification:** If more than two failures occur, the robot moves to a lookup point. The RGB-D camera re-identifies the box and machined parts, correcting orientation information.
6. **Cycle Management:** After successfully completing five operational cycles, the system triggers a return to the lookup point. This periodic lookup adapts the system to potential changes in the scene within the box.

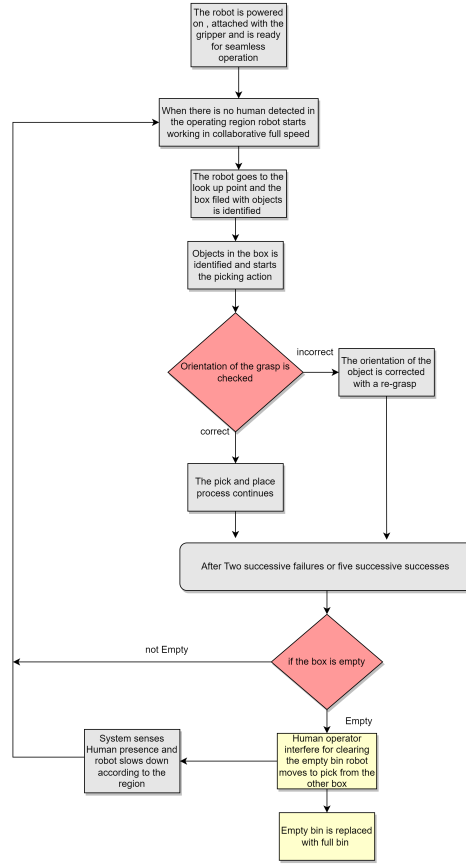


Figure 5: Flow chart of the Standard Operating Procedure

7. Human Intervention for Bin Replacement: When the first box is empty, the robot sends a signal for human intervention. A human operator replaces the empty box with a filled one.

8. Placement and Conveyor Belt: The robot continues picking from the second box and places machined parts on a conveyor belt. The robot changes its orientation to keep the time and avoid singularity errors during pick and place from the second box.

9. Specific Placement: The robot is programmed to place objects precisely in a predefined location. The magnetic gripper ensures a controlled release, preventing unintended drops.

10. End of Cycle: The operational cycle completes within the designated 28-second timeframe.

The bin-picking robot system integrates state-of-the-art components to deliver a reliable and precise solution for the automated handling of machined parts. The combination of a magnetic gripper, RGB-D camera, and Safe Human Detection sensors ensures efficiency, safety, and adaptability. The system's ability to manage potential failures and

periodic lookups contributes to its overall robustness and reliability in a manufacturing environment. Regular maintenance and adherence to safety protocols are crucial for optimal performance.

3.1.3 Control Structure

In this system, there are two computers, one would be processing the AI inputs and then sending signals to the other computer which would control the robot with the software. The AI computer after recognizing the object would send the coordinates of the detected object to the control computer. The calculation of the forward and inverse kinematics is done inside the second computer in this case, and the optimal way of approaching the workpiece is found out using this. The AI computer also sends information about the force to be used by the end effector to pick the object according to the object identified. The Robot-Human interface allows feedback from the operator as well as helps the robot to send signals about its actions. The number of successful picks and the successful failures are being monitored for this particular application to trigger the movement of the robot to look up point.

The orientation of the pick is monitored, if the orientation is not as expected, then the robot puts back the workpiece. If there are two failures in pick, and then the robot puts back the workpiece the robot has to go to the look up point , There are control triggers from the AI system in these two scenarios

The control unit would have the state and orientation of the robot which is being updated in every one millisecond by the heartbeat signals. There are passive safety controls like collision detection and safe torque ON when the robot detects anomalies in the voltage behavior and the external forces. The Safe Human Detection sensor senses the human or any obstacle inside its perimeter and then the robot slows down or the robot stops according to the region where the human is detected the Safe Human Detection sensor is connected to the Digital input of the system. As soon as there is an obstacle in the region the Safe Human Detection sensor sends the signal high into the Digital inputs and then the processor first assesses the state of the robot and then asks the robot to slow down or to stop or to be in stop mode according to the state of the robot. The monitoring of the robot limits cartesian space and the joint space is continuously happening. This is also limited according to the present ISO standards and

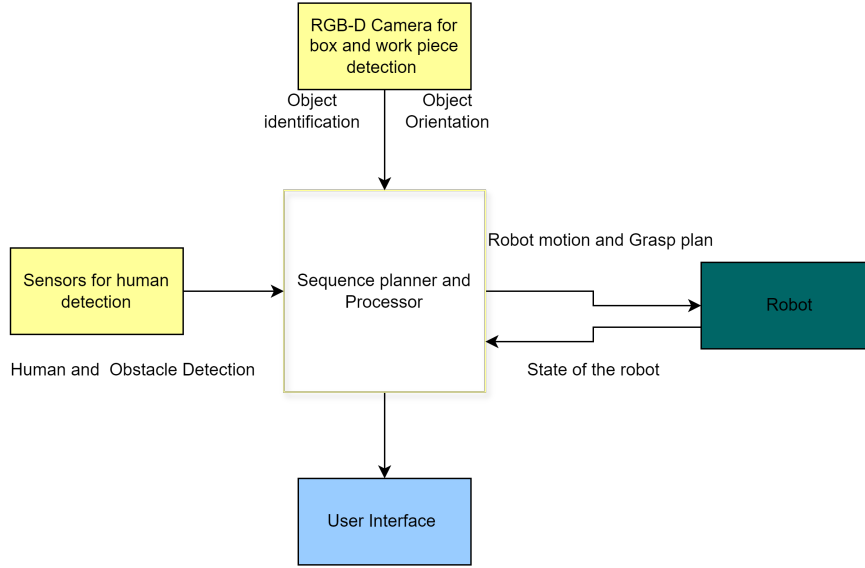


Figure 6: Control Sequence:Inputs from sensors,processing and output

technical specifications for the safe operation and integration of robots.

3.2 Risk Assessment Methods and Tools

The ISO/TS 15066 or EN ISO10218-1:2011 and ISO 10218-2:2011 do not specifically dictate a special risk assessment method. ISO 12100:2010 suggests a task-based risk assessment methodology to determine the performance level required for the safety functions.

To enhance the system and process safety there are many techniques available such as FTA, ETA, HAZOP, FMEA STPA, etc.

Fault Tree Analysis (FTA) is a top-down risk analysis method that examines an entire system by breaking it down to components and analyzing the conditions that may lead to a predefined problem. This problem is called the top event. Defining the top event, decomposing the system into individual components. Identification of relevant events and conditions, and establishing the logical relationship are the main steps of FTA. FTA allows quantitative and qualitative analysis. The result of the analysis is documented using the Fault Tree diagram, illustrating the logical relationship and the need for improvements in the system safety measures.

Hazard and Operability Studies (HAZOP Studies) are a risk analysis technique for a defined system, identifying potential risk in operating and maintaining the system and

external risk sources like environmental hazards that can trigger risk for the system itself.

Then we have risk assessment techniques like Failure mode effects (FMEA) which break down the system and analyze the risk of each of the components or each step of the process.

System Theory Process Analysis (STPA) is a comparatively new risk assessment tool that can be used to study the control structure of the system, how the interactions between the components are processed, and analyze if there are unsafe control actions. FMEA and STPA are discussed further in the next subsections as we are doing our risk analysis of the AI-integrated robotics system using an extension of FMEA with STPA

3.2.1 Failure Mode Effects Analysis

FMEA is an engineering analysis method to be used to analyze the system from bottom to up for the design of new systems, new processes, new software, and later to improve production and design. It is useful to evaluate failures that may occur, the effect of the failures on the system, and mitigation strategies. EN IEC 60812:2018 is a standard that specifies how to use this risk assessment tool effectively in the industry. This standard illustrates use of FMEA in different scenarios.

FMEA applies to hardware, software, and human action interface between humans and hardware or software in any combination. This could be used in any stage of the lifecycle of a product, process, or system. It is based on the reliability theory.

The four stages of FMEA are preparation, risk identification, risk assessment, and risk reduction. The preparation stage is where the boundary conditions are explained the process flow chart is drawn and most importantly the decisions about what should be included and excluded are considered. Risk identification starts by identifying and describing the function of relevant steps in the process and each item's potential failure modes. Risk assessment could be done by finding the RPN and estimating the likelihood of failure. Risk reduction is the step where corrective actions are developed and implemented.[30]

The analysis starts from the lowest component and proceeds up to the failure effect of the overall system. A failure effect of a lower component can be a failure mode for a component at a higher level. The potential problems in the design and development of

the system can be studied.

There are different applications for the FMEA considered such as the software FMEA and the process FMEA. The software FMEA can be used on embedded real-time systems. Software FMEA is a very important failure analysis method since the software does not fail randomly and all the software failures are systematic failures, due to wrong specifications or requirements of the function.

In this analysis, each failure mode portrays the potential failure of the product or system and these failure modes' cause of the failure and the effect of the failure are entered in the spreadsheet called the FMEA table.

In this research, we use process FMEA to analyze the process, where the robot is used to pick and place the workpiece using integrated artificial intelligence. We break down the process into steps and substeps and do the analysis

The integration of Artificial intelligence in robotics introduces a paradigm shift that brings several risks into the scenario. We use FMEA as our risk assessment tool and not other tools like HAZOP, Fault Tree Analysis, Event Tree Analysis, or Common Cause Failure Analysis because of its unique features.

FMEA is useful in identifying causal factors such as interface errors, Hardware and software interaction, and software defects in the usage of improper protocols or even the architectural mistakes in the software of the system.

FMEA allows the breakdown of the system into constituent parts so that the identification of risks is possible at the component level. In the context of AI in robotics, it is essential to break the system into constituent parts as there are many intricate algorithms and dynamic decision-making processes in the control system.

The proactive nature of FMEA also was a reason for us to choose this risk assessment tool. The other risk assessment tools use the reactive method after the occurrence of the risk. This is suitable for dynamic environments and real-time decision-making.

FMEA allows quantitative and qualitative analysis of risk. The severity, occurrence, and detection of the failure modes contribute to the quantitative analysis of the risk while the team's expertise and experience could add qualitative insights.

FMEA allows cross-disciplinary collaboration, experts from robotics, Artificial Intelligence experts, other engineers, and domain specialists. This allows views from different angles to a problem and could lead to more robust risk reduction. The

algorithmic biases uncertainties in learning and adaptability to complex situations can be tailored using FMEA.

FMEA encourages the thorough documentation of the risk assessment processes. The documentation is invaluable to trace back the actions taken. This documentation is also essential for the certification processes. In the rapidly growing field of AI, having a well-documented risk assessment methodology is essential for presenting in due diligence of the AI-based control systems.

FMEA deals with the known potential failures and introduces new countermeasures to mitigate the negative effects., it is hard to identify all risks in the early phase of the development using FMEA. Especially for intelligent control systems that may make their own decisions.

3.2.2 System Theoretic Process Analysis

System Theoretic Process Analysis-(STPA) is a top-down proactive method for analysis that is based on the System Theoretic Accident Model and processes. This is a relatively new method of risk analysis that is beyond directly related to failure events or component failures but to more complex processes and unsafe interactions among system components. This analysis is mentioned in ISO 21448:2022 which addresses functional insufficiencies of systems.

STPA uses a model of the system that consist of a functional diagram of the control system not really the physical parts in the control system but how the interactions are being worked.

STPA also includes Planning, which means defining the purpose of the analysis including the definition of the system, and boundary conditions. Identifying system losses and hazards. The second step is creating a hierarchical diagram of the control structure. The third step in STPA is an analysis of the control actions. The next step includes describing the causal factors that can include unsafe control actions and hazards, translating uncontrolled actions into constraints on the behavior of each controller. The unique features of STPA enable the risk assessment of the control systems.

STPA defines hazard as the state of a system or condition that could lead to an accident. Since an inadequate control action can lead system to a hazardous state . STPA studies whether a control action required is provided, if the control action

provided is safe if the control action being provided in wrong time or sequence and if it is provided in inadequate timing.

In this study of the control actions we consider communication failure for example delayed failure or corrupted communication inside the system since the delivery of the control command to the component from the processor is also an important aspect of this.

The systemic approach in which the whole system is considered and the interaction between the components is analyzed rather than doing risk analysis on individual components. It helps in the holistic concept of the system and the potential vulnerabilities of the system.

STPA analyses the processes inside a system. The hazards arise from interactions of the components inside the system which allows a nuanced understanding of the system working.

The proactive approach of STPA allows the early detection and identification of potential hazards, which allows for the prevention of the hazard being embedded in the system in the development cycle itself.

The strong emphasis placed on control systems by STPA examines the working of the control system and identifies the deviation from the correct working of the system. This in turn helps in identifying enhancements in control measures.

STPA not only considers technical aspects but also takes into consideration the human factors and organizational factors that could cause hazards. This helps in a holistic approach to risk assessment which could include technical and human-related risk.

The causal analysis method employed by STPA allows to identification of root causes for the potential hazard which could help in more effective mitigation of the risk from various factors contributing to the risk.

STPA is being adapted in different fields like the automotive and aerospace industries for risk assessment in intelligent control systems. This could also be used in the field of robotics.

STPA considers hazards as a challenge in managing system controls, encompassing not just failures but also hazardous successes. This implies that even if individual components operate successfully, their combination may still result in a loss of system integrity. As a result, STPA is frequently utilized in safety-critical systems and scenarios

where the most severe outcomes need to be anticipated.

A notable limitation of STPA in risk assessment is that we cannot quantify risk using STPA for example there could not be an RPN number if we use STPA for risk analysis.

3.2.3 Risk Assessment using FMEA-STPA

The purpose of this project is to conduct risk assessment for the perception-based grasping system in robotics. The analysis aims to identify, evaluate, and address potential failure modes in the perception components, algorithms, and related processes to enhance the safety, reliability, and performance of the grasping system. This may include Assessing the Sensors, Object Recognition Algorithms, Environmental conditions, Integration with the Grasping mechanism, Communication, and Data processing and how this might affect human-robot interaction.

For risk assessment of the AI systems in robots, we use the tool FMEA with an extension of STPA, which is strongly advocated by the authors of [30]. To overcome the limitations of both the tools we use this extension. We use a brainstorming session for the effect cause analysis of the identified potential use case.

For FMEA which is a bottom-to-up risk assessment strategy we study each component in the process and for STPA which is a top-to-down risk analysis we study every interaction of the components in the system and the risk assessment according to the recommendation of the standard EN ISO 10218-2:2011 for robotics systems which suggests a structured risk assessment. Since it is difficult to find out the occurrence and detection of intelligent systems we use STPA for the controllability of occurrence and detection using.

The tool complements the shortcomings of each other. For example, FMEA is component-centric or event-focused, this can lead to dangerous success meaning even if the components are not failing the whole system working together can cause a failure leading to a hazardous event. this can be monitored using STPA which identifies the risk at the system level or according to the process.

While FMEA gives the prioritized actions for the mitigation of risks. STPA gives safety requirements to be followed based on technical and human aspects of the system. The integration of these tools could be a better way to assess the intelligent complex systems The process requires a well-experienced system engineer, robotics engineer, AI specialist,

a Product manager who decides the requirement for the system, and A Functional Safety expert as the first step of the risk assessment.

We try to quantify the severity, occurrence, and detectability on a scale of 1 to 5. We use the STPA and the definition of the control structure for identifying unsafe control actions. It is crucial to tackle a particular type of risk arising from uncertainty, as responses or inputs/feedback can be appropriately supplied at the right moment (neither prematurely nor belatedly). The combination of these methodologies the advantage is FMEA gives the potential risk of failures in process aspects and the STPA gives the potentially unsafe control actions that could influence the acting decision of the system. Thus when we quantify risk, we think the logic, if there are control actions for a particular error arising due to malfunction of the components, the occurrence rating would be less for this error, which can lead to a failure, Similarly if there is no control for a particular failure, the occurrence rating would be very high. This is similar in case of detectability also. Thus STPA is relevant in the risk assessment of the intelligent control system which give inputs for the robot actions.

However, this entails inherent uncertainty, especially when utilizing machine learning techniques. For example, when the vision system furnishes information to the control system regarding the operator's position, the algorithm may exhibit uncertainty regarding the precise position, particularly if the human is partially obstructed. This uncertainty will place restrictions on when the control system can carry out an operation. In the brainstorming session, these failures arose from uncertainties that were considered, and the possible remedies were discussed.

Failure Modes

When we do the Failure modes effect analysis for the process each step of the process is considered and the failure modes are determined. Each step of the process is then subdivided and failures are determined.

So when we consider the industrial use case for grasping it can be divided into 4 major phases which can be subdivided into steps for the FMEA, which are Object Identification Phase, the Picking Phase, the Placing Phase and the Human intervention phase.

We can divide the Object identification phase into substeps like Object detection,

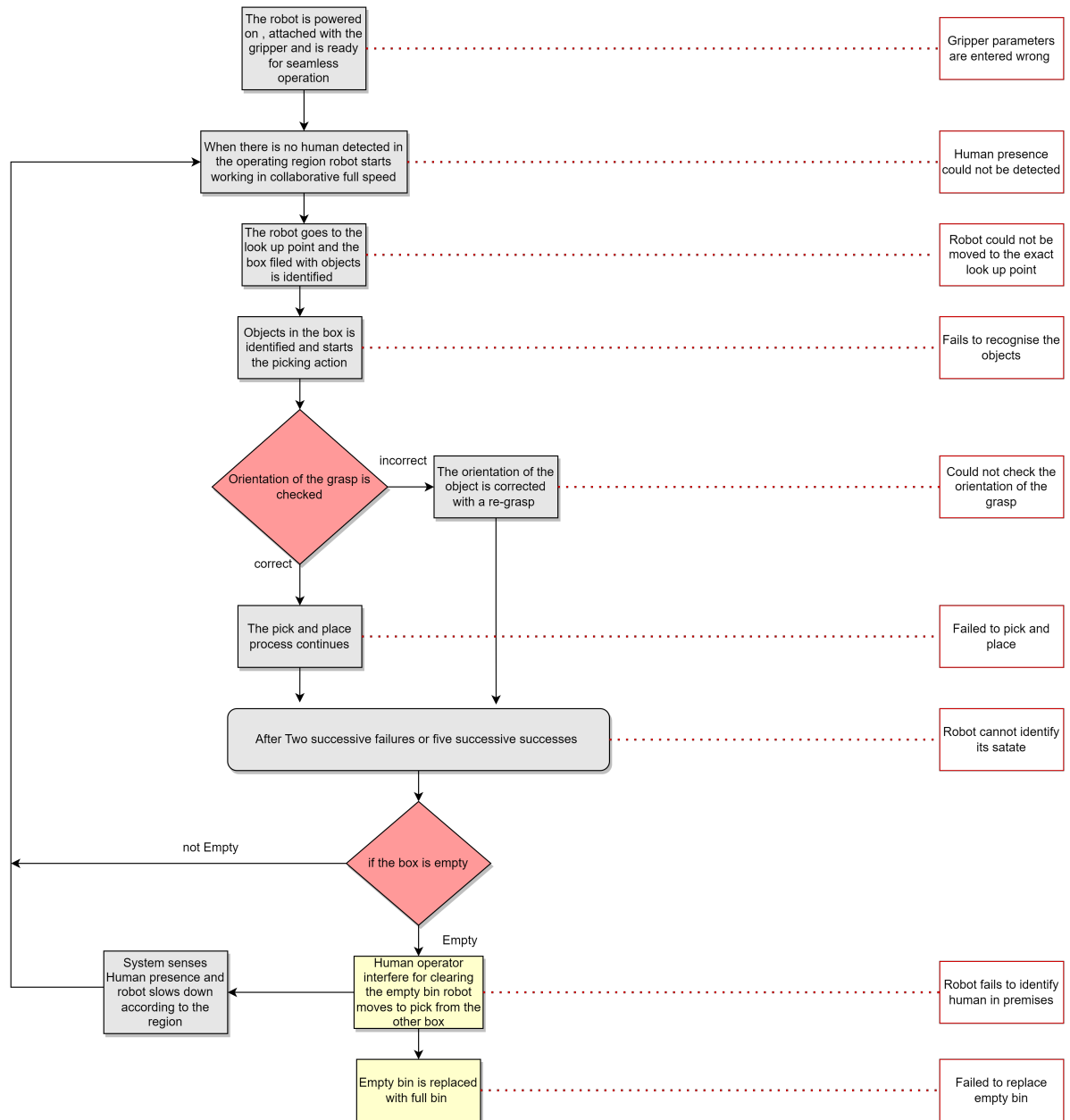


Figure 7: Failures Identified in Standard Operating Procedure

Object identification, and Grasp Estimation. The following failure modes could arise in the above phases.

1. False positives: The system detects objects even if the bin is empty.
2. False Negatives: The system is unable to detect objects inside the bin even if there are objects inside the bin.
3. Over-segmentation: Here the system takes the input and can see one object as two or more due to over-segmentation.
4. Under-segmentation: Here the system takes the camera or sensor input and sees two or more objects as one due to under-segmentation.
5. Occlusion: The system is unable to recognize the objects that are occluded due to the objects above in the box.
6. Incorrect Identification of Object: The system could incorrectly identify the objects and may damage the object due to this.
7. Incorrect Orientation Recognition of Object: The system incorrectly recognizes the object's orientation.

The second phase can be divided into substeps where the robot moves to pick the object, the robot applies the correct force to pick the object and then finally grasps. In this phase after grasping the robot checks the orientation of the object after grasp. The following failure modes could arise in the above phases.

1. Incorrect Orientation of the Gripper to Pick the Object: The gripper is oriented incorrectly, which could cause the pick of the object incorrectly.
2. Incorrect force for pick.

In the 3rd phase, the robot places the object and this can be subdivided into the moving phase where the robot holds the object and moves to the final destination, and the placing phase in which the robot places the object in the required orientation. The following failure modes could arise in the above phases.

1. Objects could be misplaced on the conveyor belt could cause damages
2. Misalignment of the robot arm during the process

Human intervention is asked, whenever the box is empty for replacement. In the Human intervention phase for replacing the bin the following failures could happen, the Safe Human Detection sensor is used for detecting human approaching and the outputs would be different as there are zones for the human to interact with the robot. The

following failure modes could arise in the above phases.

1. False negative in human identification: The system fails to identify the human presence near the system
2. Human Location Uncertainty due to Occlusion: The system identifies to detects human presence but cannot identify the exact location of the human due to Occlusion
3. Delayed Human Detection: This can cause great danger for the operator, this is a severe failure mode where there is a delay in human detection due to communication breakdown between sensors and processors or high computing time for the processor.

Control Analysis

We do a Control Analysis, to understand the communication control actions and the feedback loops between the AI system and the robot control system. The first phase of the analysis is the assessment of the clarity, reliability, and efficiency of data exchange. The analysis also includes whether there are control actions provided, not provided or provided with timing discrepancies. This assessment may help to find out the vulnerabilities of the control mechanism. The timing discrepancies mean if the control action lasts too long, if the control action stops prematurely if it happens too early, or if it happens too late.

One of the major control actions in this use case is the robot reducing its speed as soon as the system detects a human in the virtual cell where the robot is working and, when the human operator is in the second region inside the collaborative workspace that is very close to the robot, the robot stops all its movements. This action starts at the moment the human steps inside the region and lasts long till the human is moved out from this region. This is one of the active control mechanisms which is used to avoid collision between the operator and the robot. Detection, Evaluation and Identification of the box and the objects are control actions which is provided just before the pick is started

The trigger to pick and place the object. The other control action is providing the required force for holding the workpiece by the magnetic gripper. As soon as the object is identified and the robot is near the object for pick. the magnetic force should be provided for the pick. This would last till the object is placed in the correct place. Changing the states of the robot and the whole system is an action provided by the control systems from its inputs. When analyzing the system's change of state for some of

the potential failure modes we found that, there is a lack of control action thus limiting the dangerous success of the robot operation. Thus few of the control actions are recommended to be implemented for the perception-based pick-and-place robotic system. These are the implementation of fail-safe mechanisms and predictive maintenance and analysis of the sensors on the robot and the sensors in the surroundings. This is discussed in detail in the next chapter. The need for planned safety also arose from this control analysis a small description for the need for planned safety is described. In short, this control analysis also helps to analyze the interactions between the components, which helps in finding not only the failure of specific components or function of the robot but the unsafe control actions too.

Uncertain Failures

The uncertain input controls from the vision system and the sensors on the robot and the application environment reduce the trust in the system. This also acts as a barrier for the system to reach the required performance level with safety.

The uncertainty in the position of humans or any obstacle in the robot's path or in the surroundings should be reduced. The uncertain events and the events with no control preventions are identified with this process.

To implement the planned safety, this is important. An approach of planned safety and active safety together is essential for the robot system integrated with AI to mitigate unknown risks thus a deliberative safety implementation is used by the control software of the system thus the unintended signals would not be produced by the control system. Planned safety is an approach defined by the researchers [35] for integrating intelligent human-robot collaboration in Industry 4.0. This approach helps in rectifying the uncertainty from the vision system and other sensors used in the system.

3.2.4 Decision Criteria for Treatment of the failure modes

The Decision criteria for the treatment of the failure modes are based on various factors such as the severity, controllability, and probability of detection other factors like complexity integration of the treatment action to the system and the effect on the performance matrix are also considered in the decision criteria. The risk matrix is found out of the severity, occurrence, and detectability found in the table.

Here severity is assessed by taking the potential failure effect into account. We use a

five-point scale for this. The reasons and the failure mechanism is also noted down in while severity is being checked

Rating	Description	Criteria
1	Very Low	Objects not picked correctly
2	Low	compromise in cycle times, Objects not placed correctly
3	Moderate	Robot colliding, damage to the workpiece, damage to the obstacle
4	High	Damage to the robot, System shutdown
5	Very High	Injured Operator, Fatality

Table 1: Severity Rating (S) Table

While analyzing the probability of the occurrence and detection the usage of the STPA comes into the scenario where we analyse whether there are control actions provided for this potential failure mode, if the control action is provided if it is sufficient, and if it is efficient.

The probability of occurrence is taken into account considering the system's complexity, potential failure mode, and cause of failure which is also in a five-point scale. The uncertain control actions are also mentioned in this part of the table. If there are no control actions provided for this from the system the maximum occurrence rating would be assigned for this cause of failure mode.

Thus instead of quantifying the failure rate we consider whether the control action provides prevention from the failure and based on the provision of the control action and based on the efficiency of the control action we try to number the control prevention. If there is suitable control prevention to protect against potential failures, then it is given 1 and if there is no control prevention at all then it is numbered 5. The occurrence rating from 2 to 4 is assigned based on the suitability and length of the control action provided. Detectability is assessed on a five-point scale based on the complexity of the component

Rating	Criteria
1	Suitable and Efficient Control Provided
2	Suitable Control prevention for prolonged time
3	Suitable Control Prevention but wrong timing (too early or too late)
4	Suitable Control Prevention but for short time
5	No Control Prevention

Table 2: Probability of Occurrence Rating (Po) Table

and the potential cause of failure. The detectability is also measured by the ability of the processing unit to detect whether there are potential failures for the intended functionality.

If there is no detection of the cause of failure then the probability of detection value would be considered the highest. If there is suitable detection for the cause of failure then it would have the lowest rating.

Later the Risk Priority number (RPN) number is calculated by multiplying the Severity value by the Probability of occurrence times the Probability of Detection.

According to the RPN obtained for the failure modes we prioritize the mitigation methods to reduce the risks with the highest RPN. The categorisation of the RPN is explained in the next section The mitigating actions proposed should bring the risks to an acceptable level. Active safety and planned safety approaches can be used to bring down the risk to a minimum.

After the reduction of the risk, the process can be iterated to identify new potential failures and risk reduction methods. The SOP should be altered in such a way that these process failures are mitigated if needed or the system requirements should be changed in order to reduce the RPN thus the detectability and the controllability of the very severe risks can be brought down.

The major results of the risk analysis were the implementation of the control actions for the potential failure modes, including upgrades in the sensors used, and the introduction of predictive analysis algorithms for the sensors, thus detecting anomalies in the sensor

Rating	Description	Criteria
1	Almost Certain Detection	Detectable at least 90 percent of the time of failure
2	Highly likelihood of detection	Detectable at 75 percent of the times of failure
3	Moderate	Detectable at 50 percent of the times of failure
4	Low likelihood of detection	Detectable at 25 percent of the times of failure
5	No Detection	Detectable at less than 10 percent of the times of failure

Table 3: Probability of Detection Rating (Pd) Table

behavior. This is explained in the next chapter.

3.2.5 Categorisation of Risk Priority Number

The risk priority number is considered the highest when the severity X Occurrence X Detection is the highest possible number that is 125.

Based on the recommendations from the experts in the FMEA the risk priority number above 75 was considered high risk and there should be appropriate measures taken to ensure the reduction of the risk priority number.

Risk Priority numbers below 30 are considered low risk and the system can have minimal risks, these risks can still be monitored but are of lesser priority.

The Risk Priority number below 75 and above 30 falls under medium risk and this should also have measures taken to reduce the risk but it is not as priority as the high category risks.

4 Results And Discussions

The research questions answered through this thesis are what are the potential safety risks associated with AI integration in a human-robot collaborative environment? How do AI-driven systems interact with other robotic functionalities and what are the implications for overall system performance? What could be a risk analysis technique for AI-integrated robotic systems? What measures can be taken to ensure seamless integration of AI into lightweight cobots? We try to implement FMEA -STPA as a risk analysis method for the particular use case in which Artificial intelligence is integrated into the human-robot collaborative environment, which was used to find out the potential risk due to AI integration in the human-robot collaborative environment. We also compared the risks with the risks emerging from autonomous driving systems. Some of the uncertain failures were found using this data. The FMEA-STPA, risk analysis method also answers how the interaction of the AI functionalities and the other robotic functionalities takes place since STPA, analyses the interaction of different control interactions in the system. The ultimate goal of the thesis is to identify and integrate the requirements for the system and the human practices for making the lightweight robots operate beside the humans in a collaborative manner and not in the cage, by making use of the artificial intelligence technologies available. The safety requirements and measures are discussed in this chapter.

4.1 FMEA-STPA Results

The FMEA-STPA table is the result of the whole Failure mode analysis and the system theory process analysis we did on the process. This table is a summary of the FMEA-STPA which has the requirement of the process done, the potential failure mode, the effect of the failure mode, the failure mechanism, provision of control action, occurrence of the potential failure, detectability of this potential failure all noted down. Thus this table allows to determine the actual status of failure prevention and failure detection mechanisms in the control system and the improvements that can be made. . The risk priority number is calculated as the product of(severity x occurrence x detectability) is also added in this table. This table allows us to monitor the optimization actions and the assignees of the work.

This table in addition have a second RPN number which is calculated after introducing the recommendations to reduce high RPN failure modes. Thus showing the possible risk reduction by introducing methods to reduce occurrence and increase detection by the control system.

The probability of occurrence is based on the control action provided for the potential failure modes. The quantification of control action is done according to the efficiency and time of the control action.

The whole process is studied and the failures associated with AI functionalities and the post-processing of the AI signals are illustrated in this table. The process phases like object recognition, grasp estimation, pick and place, robot movement to the look-up point, and human intervention are presented. The important parts of the table is added as an annex to the thesis.

The signals sent and the control actions are studied based on the table. Where the AI functionalities are used and according to their failures. The progress of work from the assignees can be monitored and improvements can be done in an iterative method.

The justification and the improvement methods for high-risk process failures are listed below based on the table which is attached annex and the major

When the AI system was studied severe failures were found as false negatives in human detection, uncertainty in the position of the human detection, or time delay in detecting the human presence in the working space of the robot. These failures have severity of 5 in a scale of 1 to 5 for the severity. Where 5 is the most severe failure. Since the potential effect of this is fatality or injury to the operator. The potential causes for these failures are sensor malfunction, algorithmic errors communication delays, or longer processing time, The presence of contaminants on the sensor surface can also cause the problem.

For severe failure mode False negatives in human detection which has a severity of 5, in current control, we do not have a sensor fail-safe mechanism for sensor malfunction thus the probability of control of this severe failure due to this failure mode is very low, thus having a high Occurrence rating of 5. The probability of detection for this is also low, thus it also has a very high rating for detectability as 5. Thus when calculated the RPN number= 125 is the highest for this failure mode and its effect.

Recommendations are given for precautions to avoid the severity of this failure, to

model the behavior of the sensor, and to detect the anomalies in the behavior of these sensors, implementing a predictive maintenance algorithm for the sensors. Using redundant sensors and monitoring the inputs from these sensors are also recommended for using the sensor fusion data.

The fail-safe mechanism for the human detection sensors would stop the robot from working if the sensor is damaged or has errors until and unless the sensors are replaced or repaired. The usage of robust machine learning algorithms for predictive maintenance and sensor fusion data could improve the controllability and the detectability of sensor malfunction. The graphical user interface can give output to the user asking for sensor replacement and validation.

A second failure cause for this severe failure is the presence of contaminants on the sensor surface, which limits the sensor to detect human presence. There is no current control for this cause, and the detectability is very low for this. Usage of sensor fusion data can improve in detecting failure due to this cause.

The algorithm modeling the ideal behavior of the sensors and the abnormal behavior when contaminants are present could be used to find out the anomalies in the expected behavior of the system and then fail-safe is suggested to improve detectability. To control the accumulation of contaminants on the sensor surfaces, the sensor should be placed in protective spaces, which can prevent moisture and dust from accumulating on the surface.

The change in environmental conditions is another cause of failure for this failure mode of false negative in human detection, which does not have any current control and could not be detected using Sensor fusion data. The RPN value is calculated as 75 which is high. Recommendations given by the experts for reducing the high RPN number of this failure mode are the following.

Robust data training should be done to overcome the environmental changes, for example, the lighting conditions, the system should be taught with the identification of the obstacles in different lighting conditions should be used.

A different cause of failure for the severe failure mode of not detecting human presence in the workspace of the robot is due to the loss of calibration of sensors. The probability of detection of calibration loss of the sensor is very low and there are very less controls to prevent this from happening, thus this is also a failure cause with a high occurrence

and detection rating. making the RPN number high.

The calibration of the sensors could have errors arising from longer periods of usage.

The sensors should have auto-calibration capacity, as per their experience from trained data also, the machine learning algorithm should be able to detect problems in the calibration. Redundant sensors and sensor fusion data should also be used for this, these are the recommendations came up for reducing RPN for this cause of failure.

Another severe failure mode in the phase of Human intervention for box replacement in this process is the robot can detect the human presence but is not able to find the exact position, there is uncertainty in the human position, This has a severity of 5, the most common cause of this is occlusion of the human presence in the sensor, which has no current controls as of now, but could use the sensor fusion data to overcome the problem of occlusion. Planned safety is another approach recommended to overcome this failure mode, thus there is an understanding between the robot and the user about the next actions.

Delayed human detection is one of the most severe failure modes in this phase of the operation. Communication breakdown and delay in communication are potential failure cause for this. The recommendations for improving the control and detection for this were implementing heartbeat signals, between the sensors and the processor, thus the robot will not work if there are communication breakdowns between the components.

The heartbeat signals between the sensors, processor, and robot controller are all monitored. The communication breakdown can be easily detected by the system.

The communication failure between the system components due to delay in the processing of the signals can be controlled by Hardware acceleration using GPUs for overcoming processing delays. optimizing the algorithms is also recommended, so the processing times are lowered.

The next severe failures arose from the object recognition phase, different failure modes identified are false positives, false negatives, misidentification, scale and orientation misinterpretation of the bin and the workpiece, and wrong grasp estimation.

Mostly the effects of these failures are workpiece or robot damage or compromise in the cycle times, when one of the boundary conditions is the cycle time, the system should try not to compromise on it. The damage to robot eventually leads to a system shutdown or a more severe failure which should be prevented. The damage to the

workpieces causes loss in production. The potential causes for these failures are the RGB-D Camera malfunctioning, RGB-D camera calibration error, change in environmental conditions, Algorithmic errors like under-segmentation and over-segmentation, and Failure to understand variety of objects. When considering the severity of these failures it ranges from 1-3.

The False positives in Box detection can cause damage to the robot, which can also cause the unintended movement of the robot causing damage to the robot or even a threat to the operators thus the severity for this is 5. Potential causes for this failure could be error output from the RGB-D camera. There is no current control for unintended motion due to sensor malfunction, the detection of the RGB-D camera malfunction is also very low by the system the detection rating is very high as 5. Thus the RPN number for this mode of failure due to this cause is calculated as 125 . The recommendation from experts was the usage of the redundant sensor for object detection and the implementation of machine learning algorithms to find out the deviation from the actual sensor working and malfunctioning. thus fail-safe mode can be implemented. A second cause for this failure mode could be the calibration errors. Currently, we do not have control over the loss of calibration of the sensors due to prolonged usage and the detectability for this is also very low.

Enabling autocalibration of the RGB-D camera is one of the controls recommended for this. The usage of redundant sensors and robust algorithms to compare the inputs from these sensors The predictive analysis of the variation of the environmental conditions and the response of the sensors and the system should be studied. Implementing the machine learning algorithm based on this would be recommended to enhance the performance in variable environmental conditions. This could also enhance the detectability of the failure mode.

Another important cause for this failure mode would be a change in environmental conditions. This can be easily controlled in our system because of the robust algorithms we use and the change in the environment is easily detectable by the system.

The experts made recommendations to improve the algorithms by using more data to train, thus operations in diverse environments are possible.

False Negatives in Box Identification is a failure mode, which can impact the cycle execution times. This is not a severe failure, but this should be addressed since this is

an important boundary condition for the process.

In the object recognition phase, another severe failure was identified as object misidentification, this could lead to effects like picking two or three objects together, which may exceed the robot's payload, and can cause damage to the robot.

The over-segmentation and under-segmentation could lead to object misidentification and object scale and orientation misidentification. The severity could be in the range of 2-3 as this can cause damage to the robot if the robot tries to pick 2 or 3 objects in one grasp due to under-segmentation. This could be prevented by analyzing the connected components in the pixel and finding out if these are very big components and then implementing a corrective action or feedback to the operator thus dangerous success of pick can be avoided. Similarly for the over-segmentation, a check can be done if the connected components are very small or if the number of connected components is high. For Grasp estimation, the system should be able to calculate the pose and the coordinates of the objects identified using the inputs from the RGB-D camera.

Identifying the object also helps in calculating the grasping force that should be used. Wrong Grasp estimation is one of the failure modes identified in the object recognition phase which can have different failure effects.

This failure mode can lead to a failure effect where the workpiece may fall while the robot is in motion holding the workpiece due to less force applied which has a higher severity of around 4. This can cause damage to the workpiece and injury to a human operator, this can be controlled by picking the object in the correct orientation and force we can control this in our system and we have a good detectability for this. The efficiency of the grasp is very high in our system

To avoid this effect the scaling and orientation of the object are checked before the pick. Improvement in these algorithms is suggested for increasing the reliability of pick as the implementation of tactile sensors in the end effector. Usage of the fusion data from the RGB-D camera and the tactile sensor for understanding the orientation of the object. Thus the controllability and the detectability of this failure can be increased thus the RPN values can be reduced. In addition to the fusion data the implementation of real-time feedback and correction is also suggested for improvements in this.

Object variability may not be a problem for this application, since we have only one type of object to pick in this application, but it is important in a broader autonomous

robotic system. Training using more robust data with a variety of objects at different environmental conditions is suggested. The type of learning that should be implemented for this is under our research. Learning specifically each object needs very large data instead training with different classes of objects is more a suitable way for training recommended for improving the performance of the application.

4.2 Safe Practices and Safe System Requirements

While a robot integrated with an AI system is used for collaborative applications the following practices and system shall have these features in addition to the ISO standards and specifications for the robots and the integrated workspace for the robots. These requirements are formulated as a result of the risk assessment done on the Human-robot interaction application which was illustrated in the methodology, where the robot uses its perception to pick, place and identify objects and sense humans in the workspace. Additionally the informations gathered using the research for this is also used to formulate the requirements and practices for safe operation of the robot and integration into industrial collaborative environment.

These requirements are essential for diversifying the usage of robots for production, the robot should be adaptable to different payloads and working envelopes. To enhance the performance of the robot in collaborative applications, which is now low performance due to the complexity of the safety systems which separates and limits the human-robot collaboration. The lack of a systematic hazard assessment for HRC is also a cause for the low-level performance of collaborative applications of robots.

4.2.1 Safe Practices

The safe practices are useful to increase the acceptance of the robot by operators for collaborative operations.

1. A clear definition of the task undertaken by the robotic system and the environment is essential for the safe integration of the AI robot system into collaborative applications for the integrators. This includes the lighting conditions of the workspace, whether the environment is dynamic or static, the frequency of human intervention in the process, the type of human-robot collaboration, etc. Therefore, the most crucial step in determining the scope of system development is to precisely define the intended function of the system by developing comprehensive descriptions of the ideas, functions, and

constraints of the cognitive and judgment technologies that make up the system. Specifically, "Operator Misuse," which refers to incorrect operation and mistakes made by the robot operator while running the robot system, may also be taken into consideration, much as the consideration of driver misuse in autonomous vehicles. The definition of the task is important for determining the interaction level and collaboration strategies.

2. Ensure the operators follow the Standard Operating procedure for the operation thus maximum uncertain risks are avoided. Include the operators who are going to interact with the robot and include their concerns and inputs while formulating the Standard Operating Procedure.
3. Examine the triggering event, which is a risk action factor, to determine the sensor or controller's algorithmic constraints as well as the circumstances that could result in a safety objective violation. This is the result of a lengthy procedure that involves defining and analyzing the triggering event.
4. Ensuring robust cybersecurity to protect the robot from unauthorized use.
5. Training for the users on how to interact with the robot and tackle uncertain failures. Also educating the users about the limitations of the robot.
6. Educate users about the system's capabilities and limitations.
7. Collaborate with domain experts and other stakeholders to ensure that the decision-making process aligns with ethical standards and societal values. Seek input from experts to enhance transparency and address potential biases.
8. Proper maintenance and validation of the sensors and the other hardware in the system as a whole. Re-calibration of the robot and the sensors if there are errors. Keeping the maintenance record for the sensors and keeping a check on the errors.
9. Training the operators to make use of the planned safety and responsibility building

4.2.2 Safe System Requirements

These requirements would enhance the safety of the AI systems in robots and would be helpful to build trust for the users.

1. Integrated redundant and diverse sensors for taking in different measurements, to find out the presence of humans and obstacles, to monitor and diagnose the status of the robot itself. In a broader view when we have to roll out real collaborative applications using the robot the sensor and processors must be also capable of supporting this.

Instead of only detecting and sending the signal to digital inputs, A processing algorithm should work to process the sensor inputs and accelerate the collaborative applications of the robot.

2. Online motion planning according to the inputs from Sensor Data after estimating the confidence of all the sensor data. Also, use of the reflex in motion planning to avoid collision and even after a collision to limit the damage to the robot and the object. This would also help in attaining the time limit of the cycle execution.

3. Provide visual feedback to users about the robot's perception of the environment, including detected objects, obstacles, and the robot's planned actions. Use visual indicators to communicate the robot's intent and decision logic. Integrating this into the user interface could improve the user experience and safety

4. Ability execute planned safety[35] by the system. There is reactive safety in robots integrated with artificial intelligence, which is common in automation systems. where the robots stop or slow down when human is in the safe region or restricted region which may lead to larger cycle times. If the autonomous robotic system wants human intervention, for example, if an object falls down from the robot's end effector while moving, it might have to be picked up by a human. The robot can send signals for the intervention and when the operator is in the perimeter, the robot can throw visual feedback, that it is going to continue executing the cycle but in a different orientation or motion, if the operator agrees on this he can give input from his side and then the robot can execute the replanned action and the operator can put the piece in place. Thus the cycle time is not compromised and the robot and the operator can make their decisions and communicate with each other and the process is done flawlessly. Planned safety is an alternative for active safety which has less concurrency and needs more complex monitoring system.

5. The addition of visual cues for communication is a requirement for the system to execute planned safety. The introduction of a Universal visual cue system for robot-human communication would greatly accelerate the planned safety for collaborative applications.

6. A system for reporting errors or uncertainties in the robot's decision-making. Provision of visual signals when the robot encounters situations it cannot handle or where its confidence is low and needs human intervention for the further process.

Integrate human-in-the-loop systems that enable human operators to intervene or override the robot's decisions when necessary. Communicate when the system is operating autonomously and when human intervention is required. A generic feedback system should be developed for this.

7. Implement interactive communication mechanisms that allow users to query the robot about its decisions. Enable the robot to provide additional information or clarification upon user request.
8. Definition of the confidence levels at which the robot takes certain actions. Clearly explain how the robot's confidence influences its behavior.
9. Redundant task planners and task evaluators should be present in the system for the task planning and verification of the task according to the inputs of AI functionalities.
10. Establish a feedback loop for continuous improvement based on user feedback and real-world performance. Use user feedback to identify areas for improvement in transparency and decision-making.
11. The system should be able to periodically assess the dynamic surroundings to evaluate the change in the dynamic environment. This helps in adapting to the changes that may occur.
12. The processing and control unit must possess robust computational power to efficiently process the extensive data from vision sensors. This should ensure the uninterrupted and timely execution of the actions based on the real-time data.
13. Integrated predictive maintenance in the system which allows the system to notify the operator about the status of the system components including the robot, sensors, and output devices based on the data for vision and other sensors according to the historical performance data, calibration data, Image quality, pattern recognition performance, dust and contamination, system alignment and performance. For robot, output vibrations, temperature, friction parameters current used, and the torque in the motor, as well as accuracy and precision of the robot operations.
14. The addition of robotic skin and extra sensing modules for the high-load robot is a suggestion from researchers, for making the high payload industrial robots collaborative.
15. An algorithm which can assess the changes in the environment , for assessing the criticality of the changes in this output. Quantification of the risk using this algorithms. Thus not only uncertain hardware failures are taken into account, machine learning

algorithmic failures can also be monitored.

In situations where the human-robot collaboration is very low and the autonomy is low, where the robot will do only repetitive tasks in a cage, operators have substantial control, and operations are comparatively safe. The responsibility often falls on the operator to promptly control the system and correct any errors. In high-level autonomy for robots which are now being introduced into the market, where both perception tasks and safety responsibilities rest primarily with the robotic system, the ability of the perception system to self-identify errors becomes paramount. These requirements for the system and practices establish a standardized protocol and enhance common terminology in robotic system function and safety aspects to promote interoperability in the field of collaborative robotics.

5 Conclusion and Future Prospects

The integration of artificial intelligence (AI) into robotics has significantly contributed to the automation of industries, playing a pivotal role in the 4th Industrial Revolution. This transformative technology is rapidly advancing toward fully autonomous robots, yet the standards governing their operations are still evolving. A comprehensive examination of EN ISO 10218-1:2011, EN ISO 10218-2:2011 and ISO/TS 15066 revealed a gap in specifying do's and don'ts and system requirements for AI-powered robots.

In response to this gap, our thesis delved into the identification and quantification of risks associated with AI failures in robotics, utilizing Failure Mode Effect Analysis (FMEA) as a powerful tool. Following guidelines from EN IEC 60812, we assessed the severity, occurrence, and detectability of different failure modes, calculating Risk Priority Numbers (RPNs) to prioritize and address potential risks. We also studied the interaction of AI functionalities with other control functionalities of the robot.

In the system relying solely on FMEA could be challenging. At the same time, System Theory Process Analysis is a tool mentioned in ISO 21448 that offers a top-down approach for the structural analysis of control commands which supports a comprehensive assessment of the intelligent system. The integration of STPA in an industrial use case may be challenging thus we did the FMEA- STPA which is an extension of FMEA using system theory process analysis, on one of the industrial use cases, where the robot picks and places machined parts from a bin to a conveyor. The perception system integrated into the robot, and the interactions inside the system between each component are also studied for this risk analysis. The individual failures of the system components are studied, also the failure of the system as a whole due to the dangerous success of the individual components contributing to the system working is studied in the risk assessment

We found out the potential failures in the integration of artificial intelligence in robots given perception and sensing. The severe failures could be false negatives from human detection, and the highly occurring failure would be the wrong orientation of the object in grasp. These failures could lead to risks. The risks can be from damage in the parts being handled to the fatality of the operator. Some of the failures could lead to damage to the robots themselves, which would lead to system shutdown and loss due to the shutdown.

Proposals were made for improved hardware specifications, emphasizing sensors, cameras, and processors. Architectural enhancements for AI algorithms were recommended to improve overall system reliability.

Guidelines were outlined for the development of monitoring algorithms to enhance real-time risk assessment and intervention. Recommendations were made for safe human interventions, including proper robot and AI functionalities documentation to prevent misuse. A generic safety framework for AI in robotics was formulated, serving as a foundational guide for safe practices.

We also recommend collaboration with the AI community to establish industry standards for risk assessment, the need for shared practices is also one of our recommendations for a future safety standard for AI-driven robots. Thus a collaborative effort is essential for addressing the multidimensional challenges.

As mentioned in the beginning of the thesis, it tries to address the following research questions

- 1) What are the potential safety risks associated with AI-Integration in human-robot collaborative environments? This question is answered by breaking down the perception based grasping as 4 major phases and then substeps and identifying the potential failure modes from this. This is identified in the Chapter 3 Methodology and mentioned under the heading Failure Modes, section 3.2.3. Risk Assessment using FMEA-STPA.
- 2) How do AI-driven systems interact with other robotic functionalities, and what are the implications for overall system performance? This is answered under the heading Control Analysis, in the section 3.2.3 in the chapter 3, Methodology.
- 3) What could be a risk analysis technique for AI-integrated robotic systems? This is also mentioned in the chapter Methodology, how the risk assessment of the AI integrated human robot collaboration can be done and the results of this risk assessment method is mentioned in Chapter 4, Results and Discussions.
- 4) What measures can be taken to ensure seamless integration of AI into lightweight cobots? This is answered in Chapter 4, Results and Discussions and this is derived from the FMEA-STPA table given in annex as well as from the literature review.

As FMEA-STPA is a qualitative analysis method and in this thesis we researched on the failures from the sensors and if it went unidentified what could be the RPN and what measures should be done to reduce the risk priority number. For the quantification of

the uncertainty of risks from AI algorithms while integrating in human robot collaborative environments, development of a machine learning algorithm as mentioned in the Results, which can assess the risks in the output according to the changes in the input is the next important step to be done.

Acknowledging the limitations of the study, we advocate for the continual refinement of risk assessments on the complex AI systems with dynamic data and potential biases with unforeseen risk using additional tools similar to SOTIF analysis used in the Automotive Industry for diverse AI functionalities. Achieving transparent decision-making in AI-integrated robots improves the trust in the AI-integrated robotic systems involves providing clarity about the robot’s actions and the reasoning behind those actions. Implementing planned safety complementing active safety is also one of the suggestions from the result of the thesis. This can be the future development in the field of safety and building trust for AI integration in robotics.

The focus of our FMEA-STPA was on perception-based grasping, providing a starting point for a broader safety framework. Future endeavors could extend this analysis to different scenarios, AI functionalities (e.g., NLP and perception, gesture control), and robot types, fostering an iterative process to enhance safety comprehensively. We envision that our research lays the foundation for the safer deployment of AI and robotics not only in industrial settings but also in Autonomous Guided Vehicles and humanoid robots. This contribution marks a crucial step towards ensuring the responsible and secure integration of evolving technologies into our daily lives.

References

- [1] Aldinhas Ferreira, Maria Isabel; Fletcher, Sarah R. (2022): The 21st Century Industrial Robot: When Tools Become Collaborators. Cham: Springer International Publishing (81).
- [2] International Organization for Standardization. (2011). Robots and robotic devices - Safety requirements for industrial robots - Part 1: Robots (ISO 10218-1:2011). ISO.
- [3] George Michalos, Panagiotis Karagiannis, Nikos Dimitropoulos, Dionisis Andronas, and Sotiris Makris(2022): Human Robot Collaboration in Industrial Environments: The 21st Century Industrial Robot: When Tools Become Collaborators.(pp 17-39) Cham: Springer International Publishing (81).
- [4] Alexandra Dobrokvashina, Shifa Sulaiman, Aidar Zagirov,Elvira Chebotareva, Hsia Kuo-Hsien,Evgeni Magid (uuuu-uuuu): Human Robot Interaction in Collaborative Manufacturing Scenarios: Prospective Cases: IEEE (2022 International Siberian Conference on Control and Communications (SIBCON)).
- [5] Murino, Teresa; Di Nardo, Mario; Pollastro, Daniele; Berx, Nicole; Di Francia, Angela; Decré, Wilm et al. (2023): Exploring a cobot risk assessment approach combining FMEA and PRAT. In Quality & Reliability Eng 39 (3), pp. 706–731. DOI: 10.1002/qre.3252 .
- [6] Matheson, Eloise; Minto, Riccardo; Zampieri, Emanuele G. G.; Faccio, Maurizio; Rosati, Giulio (2019): Human–Robot Collaboration in Manufacturing Applications: A Review. In Robotics 8 (4), p. 100. DOI: 10.3390/robotics8040100.
- [7] Valori, Marcello; Scibilia, Adriano; Fassi, Irene; Saenz, José; Behrens, Roland; Herbster, Sebastian et al. (2021): Validating Safety in Human–Robot Collaboration: Standards and New Perspectives. In Robotics 10 (2), p. 65. DOI: 10.3390/robotics10020065.
- [8] Weng, Yueh-Hsuan; Chen, Chien-Hsun; Sun, Chuen-Tsai (2009): Toward the Human–Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots. In Int J of Soc Robotics 1 (4), pp. 267–282. DOI: 10.1007/s12369-009-0019-1

- [9] Hamid Tabani, Francisco J. Cazorla, Guillem Bernat (Ed.) (2019): Assessing the Adherence of an Industrial Autonomous Driving Framework to ISO 26262 Software Guidelines. DAC '19: The 56th Annual Design Automation Conference 2019. Las Vegas NV USA, 02 06 2019 06 06 2019. New York, NY, USA: ACM.
- [10] International Organization for Standardization. (2011). Robots and robotic devices - Safety requirements for industrial robots - Part 2: Robots (ISO 10218-2:2011).
- [11] International Organization for Standardization. (2016). Robots and robotic devices — Collaborative robots (ISO/TS 15066).
- [12] International Organization for Standardization/International Electrotechnical Commission. (2023). Information technology — Artificial intelligence — Guidance on risk management (ISO/IEC 23894).
- [13] Estonian Centre for Standardisation. (2018). EVS-EN IEC 60812:2018 - Analysis techniques for system reliability - Procedure for failure mode and effects analysis (FMEA) (ISO/IEC 60812:2006, IDT). Tallinn, Estonia: EVS.
- [14] Li, Jamy; Chignell, Mark (2022): FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. In *AI Ethics* 2 (4), pp. 837–850. DOI: 10.1007/s43681-022-00145-9.
- [15] Du, Guoguang; Wang, Kai; Lian, Shiguo; Zhao, Kaiyong (2021): Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. In *Artif Intell Rev* 54 (3), pp. 1677–1734. DOI: 10.1007/s10462-020-09888-5.
- [16] Bonci, Andrea; Cen Cheng, Pangcheng David; Indri, Marina; Nabissi, Giacomo; Sibona, Fiorella (2021): Human-Robot Perception in Industrial Environments: A Survey. In *Sensors* (Basel, Switzerland) 21 (5). DOI: 10.3390/s21051571.
- [17] Brijacak, Inka; Yahyanejad, Saeed; Reiterer, Bernhard; Hofbaur, Michael (2017): Toward Safe Perception in Human-Robot Interaction.
- [18] Kaipa, Krishnanand N.; Kankanhalli-Nagendra, Akshaya S.; Kumbla, Nithyananda B.; Shriyam, Shaurya; Thevendria-Karthic, Srudeep Somnaath; Marvel, Jeremy A.;

- Gupta, Satyandra K. (2016): Addressing perception uncertainty induced failure modes in robotic bin-picking. In *Robotics and Computer-Integrated Manufacturing* 42, pp. 17–38. DOI: 10.1016/j.rcim.2016.05.002.
- [19] K. Hsiao, M. Ciocarlie, P. Brook (2011): Bayesian grasp planning, in: *ICRAWorkshop on Mobile Manipulation: Integrating Perception and Manipulation*.
- [20] Burgess-Limerick, Ben; Lehnert, Chris; Leitner, Jurgen; Corke, Peter (2022): *DGBench: An Open-Source, Reproducible Benchmark for Dynamic Grasping*.
- [21] Sun, Yu; Falco, Joe; Roa, Maximo A.; Calli, Berk (2022): Research Challenges and Progress in Robotic Grasping and Manipulation Competitions. In *IEEE Robot. Autom. Lett.* 7 (2), pp. 874–881. DOI: 10.1109/LRA.2021.3129134.
- [22] Y. Bekiroglu et al., "Benchmarking Protocol for Grasp Planning Algorithms," in *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 315–322, April 2020, doi: 10.1109/LRA.2019.2956411.
- [23] Bottarel, Fabrizio; Vezzani, Giulia; Pattacini, Ugo; Natale, Lorenzo (2020): *GRASPA 1.0: GRASPA is a Robot Arm graSping Performance BenchmArk*. In *IEEE Robot. Autom. Lett.* 5 (2), pp. 836–843. DOI: 10.1109/LRA.2020.2965865.
- [24] Sun, Chen; Zhang, Ruihe; Lu, Yukun; Cui, Yaodong; Deng, Zejian; Cao, Dongpu; Khajepour, Amir (2023): Toward Ensuring Safety for Autonomous Driving Perception: Standardization Progress, Research Advances, and Perspectives. In *IEEE Trans. Intell. Transport. Syst.*, pp. 1–19. DOI: 10.1109/TITS.2023.3321309.
- [25] P. Skruch, M. Szelest, M. Dlugosz and D. Cieslar, "Safety of Perception Systems in Vehicles of High-Level Motion Automation," 2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA), Stary Smokovec, Slovakia, 2022, pp. 561–566, doi: 10.1109/ICETA57911.2022.9974838.
- [26] Choi, K. L., Kim, M. J., Kim, Y. M. (2022). On Safety Improvement through Process Establishment for SOTIF Application of Autonomous Driving Logistics Robot. *International Journal of Internet, Broadcasting and Communication*, 14(1), 209–218. <https://doi.org/10.7236/IJIBC.2022.14.1.209>

- [27] Mohseni, Sina; Pitale, Mandar; Singh, Vasu; Wang, Zhangyang (2019): Practical Solutions for Machine Learning Safety in Autonomous Vehicles
- [28] Madala, Kaushik; Krishnamoorthy, Jayalekshmi; Avalos Gonzalez, Carlos; Shivkumar, Abhishek; Solmaz, Mert (2022): Contributing Factors to Consider While Defining Acceptance Criteria and Validation Targets for Assuring SOTIF in Autonomous Vehicles. In : SAE Technical Paper Series. WCX SAE World Congress Experience, APR. 05, 2022: SAE International 400 Commonwealth Drive, Warrendale, PA, United States (SAE Technical Paper Series).
- [29] Peng, Liang; Wang, Hong; Li, Jun (2021): Uncertainty Evaluation of Object Detection Algorithms for Autonomous Vehicles. In *Automot. Innov.* 4 (3), pp. 241–252. DOI: 10.1007/s42154-021-00154-0.
- [30] Hanna, Atieh; Bengtsson, Kristofer; Larsson, Simon; Götvall, Per-Lage (2023): Risk Assessment for Intelligent and Collaborative Automation System by Combining Fmea and Stpa.
- [31] Qi, Yi; Dong, Yi; Khastgir, Siddhartha; Jennings, Paul; Zhao, Xingyu; Huang, Xiaowei (2023): STPA for Learning-Enabled Systems: A Survey and A New Practice.
- [32] STAMATIS, D. H. (2003): Failure mode and effect analysis. FMEA from theory to execution. 2nd ed., rev. and expanded. Milwaukee, Wisconsin: ASQC Quality Press.
- [33] Adriaensen, A.; Pintelon, L.; Costantino, F.; Di Gravio, G.; Patriarca, R. (2021): An STPA Safety Analysis Case Study of a Collaborative Robot Application. In *IFAC-PapersOnLine* 54 (1), pp. 534–539. DOI: 10.1016/j.ifacol.2021.08.061.
- [34] Abdulkhaleq, Asim; Wagner, Stefan; Lammering, Daniel; Boehmert, Hagen; Blueher, Pierre (2017): Using STPA in Compliance with ISO 26262 for Developing a Safe Architecture for Fully Automated Vehicles. DOI: 10.48550/arXiv.1703.03657.
- [35] Hanna, Atieh; Larsson, Simon; Götvall, Per-Lage; Bengtsson, Kristofer (2022): Deliberative safety for industrial intelligent human–robot collaboration: Regulatory challenges and solutions for taking the next step towards industry 4.0. In *Robotics*

and Computer-Integrated Manufacturing 78, p. 102386.

DOI:10.1016/j.rcim.2022.102386.

- [36] Salah, Bashir; Alnahhal, Mohammed; Ali, Mujahid (2023): Risk prioritization using a modified FMEA analysis in industry 4.0. In Journal of Engineering Research. DOI: 10.1016/j.jer.2023.07.001.
- [37] Larsson, Simon; Bengtsson, Kristofer (2022): Enabling human-robot collaboration and intelligent automation in the automotive industry: A study of stakeholder perspectives.
- [38] Liu, Hu-Chen; Liu, Long; Liu, Nan (2013): Risk evaluation approaches in failure mode and effects analysis: A literature review. In Expert Systems with Applications 40 (2), pp. 828–838. DOI: 10.1016/j.eswa.2012.08.010.
- [39] Lee, Sukhan; Lee, Moonju; Kim, Jaewoong; Yoo, Kyeongdae; Barajas, Leandro G; Menassa, Roland (2012): 3D Visual Perception System for Bin Picking in Automotive Sub-Assembly Automation. [Place of publication not identified]: IEEE.
- [40] Oh, Jong-Kyu; Lee, Sukhan; Lee, Chan-Ho (2012): Stereo vision based automation for a bin-picking solution. In International Journal of Control, Automation and Systems 10 (2), pp. 362–373. DOI: 10.1007/s12555-012-0216-9.
- [41] Sulaman, Sardar Muhammad; Beer, Armin; Felderer, Michael; Höst, Martin (2019): Comparison of the FMEA and STPA safety analysis methods—a case study. In Software Qual J 27 (1), pp. 349–387. DOI: 10.1007/s11219-017-9396-0.
- [42] Hanna, Atieh (2021): Towards intelligent and collaborative automation of automotive final assembly. Västerås: Mälardalen University (Mälardalen University Press Licentiate Theses, 305).
- [43] Wang, Xuanyu; Qi, Xudong; Wang, Ping; Yang, Jingwen (2021): Decision making framework for autonomous vehicles driving behavior in complex scenarios via hierarchical state machine. In Auton. Intell. Syst. 1 (1). DOI: 10.1007/s43684-021-00015-x.
- [44] Pulikottil, Terrin Babu; Pellegrinelli, Stefania; Pedrocchi, Nicola (2021): A software tool for human-robot shared-workspace collaboration with task precedence

constraints. In *Robotics and Computer-Integrated Manufacturing* 67, p. 102051.
DOI: 10.1016/j.rcim.2020.102051

- [45] Inam, Rafia; Raizer, Klaus; Hata, Alberto; Souza, Ricardo; Forsman, Elena; Cao, Enyu; Wang, Shaolei (2018): Risk Assessment for Human-Robot Collaboration in an automated warehouse scenario. In : 2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA). 2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA). Turin, 04/09/2018 - 07/09/2018: IEEE, pp. 743–751.
- [46] Bitsch, Friedemann.; Guiochet, Jérémie.; Skavhaug, Amund. (Eds.) (2016): Computer Safety, Reliability, and Security. 35th International Conference, SAFECOMP 2016, Trondheim, Norway, September 21-23, 2016, Proceedings. Cham: Springer International Publishing; Imprint: Springer (Lecture Notes in Computer Science, 9922).
- [47] International Organization for Standardization. (2022). Road Vehicles- Safety of the intended functionality (ISO 21448:2022). ISO.

Annex 6

[illegible]

Table 4: Process FMEA table for perception-based grasping Phase 1

Process FMEA for Perception based Grasping with STPA extension																
Item			Characteristics of Failure				Current Control						RPN Rating			
Process Phase	Process Step	Requirement	Potential Failure Mode	Potential Effects of Failure	S	Potential Cause of Failure/Unsafe Control actions	Control Prevention	O	Recommendations to reduce High Occurrence	Reduced O	Control Detection	D	Recommendations to reduce High D rating	Reduced D	RPN1	RPN2
5	Human Intervention for replacing boxes															
5.1	Robot slows down detecting human in maximum space	Robot shall slow down as soon as it detects human presence in maximum space	False Negative	Injury or Fatality of Human Operator	5	Sensor malfunction	No current control Prevention	5	Implement Fail-Safe mechanism for human detection sensors and Planned Safety rather than proactive safety in close human robot collaboration	2	No current control Detection	5	Usage of redundant sensor data and machine learning algorithms to detect anomalies in sensor behaviour	2	125	20
									Real time operating system allowing inputs from redundant sensor inputs.	2	No current control Detection	5	implementation of heartbeat signals to monitor communication break down.	2	125	30
									Implementing fail-safe when proper sensor inputs are not possible.	2	No current control detection	5	Modelling abnormal behavior of the sensor when contaminants are present thus, detection is possible.	2	125	30

Table 5: Process FMEA table for perception-based grasping: Phase 5.1

Process FMEA for Perception based Grasping with STPA extension															
Item		Characteristics of Failure					Current Control					RPN Rating			
Process Phase	Process Step	Potential Failure Mode	Potential Effects of Failure	S	Potential Cause of Failure/Unsafe Control actions	Control Prevention	O	Recommendations to reduce High Occurrence	Reduced O	Control Detection	D	Recommendations to reduce High D rating	Reduced D	RPN1	RPN2
5	Human Intervention for replacing boxes														
5.3	Robot stops detecting human in protective space	False Negative	Injury or Fatality of Human Operator	5	Sensor malfunction	No current control Prevention	5	Implement Fail-Safe mechanism for human detection sensors and Planned Safety rather than proactive safety in close human robot collaboration	2	No current control Detection	5	Usage of redundant sensor data and machine learning algorithms to detect anomalies in sensor behaviour	2	125	20