Deliverable 5.1

Assessmentkatalog Sammlung

(Assessment Catalogue Collection)

Dokument Version	:	Final
Abgabe	:	31/01/2021
Dissemination Level	:	Öffentlich
Beitrag zu	:	WP5
Dokument Inhaber	:	UNIVIE
Dokument Name	:	ComplAI-D5.1 Assessmentkatalog Sammlung
Revision	:	1.0
Projektakronym	:	compl@i
Projekttitel	:	Collaborative Model-Based Process Assessment for trustworthy AI in Robotic Platforms
Grant Agreement n.	:	33755860
Call	:	IDEEN LAB 4.0 (2019)
Projektdauer	:	12 Monate, beginnend mit 01/02/2020
Website	:	https://complai.innovation-laboratory.org/

Revision History

REVISION	DATE	INVOLVED PARTNERS	DESCRIPTION
0.1	01/05/2020	BOC	Initial, content collection
0.2	01/06/2020	BOC, JR, UNVIE, JKU	Collection of security and safety questions
0.3	01/07/2020	BOC, JR, UNVIE, JKU	Extension, conclusion, abstract
0.4	21/12/2020	BOC, JR, UNVIE, JKU	Integration of ethic criteria catalogue
0.5	05/01/2021	UNVIE, JKU	Integration of law catalogue
1.0	31/01/2021	BOC	Transformation into public deliverable

List of Contributors:

Crompton Laura (UNIVIE), Funk Michael (UNIVIE)

Breiling Benjamin (JR), Dieber Bernhard (JR)

Schumann Stefan (JKU), Bruckmüller Karin (JKU)

Sumereder Anna (BOC), Utz Wilfrid (BOC), Woitsch Robert (BOC)

List of Reviewers:

Crompton Laura (UNIVIE)

Schumann Stefan (JKU)

Kurzfassung

Dieses Deliverable zielt auf die Erstellung von Kriterienkatalogen zum Assessment von ethischen, rechtlichen und sicherheitsrelevanten Fragestellungen ab.

Das Ziel ist nicht die Erstellung eines vollständigen Kriterien- und Assessment- Fragenkataloges. Es ist viel mehr die Identifizierung einer ausreichenden Menge an Fragen, um die Anwendbarkeit – in Synchronisation mit Arbeitspaket 4 – zu prüfen, sowie die Identifizierung von Hindernissen und Forschungsfragen. Basierend auf konkreten Beispielen werden potenzielle Risiken für Geschäftsprozesse mit Hilfe einer interdisziplinären Arbeitsweise definiert. Da die Transformation der Assessmentkataloge in formal ausführbare Metamodelle nicht trivial ist, liegt der Fokus darauf, ein repräsentatives Set von Kriterien für jede Dimension zu finden, sodass diese Sets mit dem Umfang des Projektes abgestimmt sind. Folgende Assessmentkataloge werden aufgeführt (1) Sicherheit, (2) Gefahrlosigkeit, (3) Recht, und (4) Ethik.

Executive Summary

This deliverable is concerned with the creation of a criteria catalogues for ethical, legal, security and safety questions.

The aim is not to create complete catalogues of criteria and assessment questions, it is rather about identifying a sufficient number of questions for examining the applicability – in synchronisation with work package 4 – as well as for identifying challenges and research questions. Based on concrete samples, potential risks for business processes are defined in an interdisciplinary way. As the transformation of the catalogue in formal executable meta models is not a trivial process, a representative set of criteria for each dimension was used so that those fit to the scope of this project. Following assessment catalogues are covered (1) security, (2) safety, (3) legal aspects, and (4) ethics.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
DX.Y	Deliverable of workpackage X with number Y of complAI
IEEE	Institute of Electrical and Electronics Engineers, widely known publishing and standardisation association
ISO	International Organisation of Standards
JKU	Partner Organisation Johannes Kepler Universität Linz

1. Introduction

This document is concerned with creating a criteria catalogues for ethical, legal, security and safety questions. It is based on the specified application scenarios defined in D3.1.

The aim is not to create complete catalogues of criteria and assessment questions, it is rather about identifying a sufficient number of questions for examining the applicability as well as for identifying challenges and research questions. In synchronisation with work package 4 it is ensured that the found criteria can be applied in models. The creation of the assessment catalogues was based on Rawls concept of reflective equilibrium (Daniels, 2020) that applies a deliberative process in which ethics, law, safety and security agree and adapt dynamically for underlying challenges and requirements. Based on concrete samples, potential risks for business process are defined in an interdisciplinary way. Due to varying terminologies and approaches of the individual disciples there is a risk of misunderstandings that is reduced by a common terminology. As the transformation of the catalogue in formal executable meta models is not a trivial process, a representative set of criteria for each dimension was used so that those fit to the scope of this project.

Following assessment catalogues are covered:

1. Security – The security assessment questions should ensure protection against unauthorized access, unauthorized operation, unauthorized usage, unauthorized changes, eavesdropping, publication of information to unauthorized third parties and denial of service attacks, as well as react to security violations by notifying reporting and automatically taking timely corrective actions.
2. Safety – The safety assessment questions should ensure the compliance with ISO 10218-2:20111 and ISO 13482:20142.
3. Legal Aspects – The assessment questions should ensure the compliance with regulations and laws – in specific focusing on criminal law.
4. Ethics – The ethic assessment questions should ensure the compliance with ethical behaviour and interaction.

1.1 Relation to Work Package

The deliverable at hand is the first part of work package 5, which is composed of:

1. D5.1 Assessment Catalogue Collection
2. D5.2 Final Publication

It emerged based on a close collaboration with legal, ethical and robotic experts.

1.2 Document Structure

Above, executive summaries in German as well as in English are provided in order to get an overview of the deliverable at hand.

This deliverable is composed of an introduction, in which an overview is provided. This is followed by assessment questions for security (Chapter Assessment Catalogue for Security), safety (Chapter Assessment Catalogue for Safety), legal aspects (Chapter Assessment Catalogue for Legal Aspects) and ethics (Chapter Assessment Catalogue for Ethics). The results are summarized in a short conclusion in Chapter Conclusion.

[1] ISO 10218-2:2011 (2011). ISO 10218-2:2011 “Robots and robotic devices — Safety requirements for industrial robots — Part 2: Robot systems and integration”. Retrieved from https://www.iso.org/standard/41571.html (last visited on 18/11/2020).

[2] ISO 13482:2014 (2014). ISO 13482:2014 “Robots and robotic devices — Safety requirements for personal care robots”. Retrieved from https://www.iso.org/standard/53820.html (last visited on 18/11/2020).

2. Assessment Catalogue for Security

Table 1 summarizes the security related assessment questions grouped by key aspects.

In specific, seven key aspects that are essential for security related questions were identified. Those should ensure protection against unauthorized access, unauthorized operation, unauthorized usage, unauthorized changes, eavesdropping, publication of information to unauthorized third parties and denial of service attacks, as well as react to security violations by notifying reporting and automatically taking timely corrective actions.

Furthermore, there can be a differentiation of four security levels. Not only the questions, but also the answers might differ throughout the various security levels. The first security level deals with the prevention of unauthorized disclosure of information via eavesdropping or casual exposure. The second security level is related to the prevention of unauthorized disclosure of information to an entity actively searching for it using simple means with low resources, generic skills and low motivation. The third security level is about the prevention of unauthorized disclosure of information to an entity actively searching for it using sophisticated means with moderate resources, identification and authentication control system specific skills and moderate motivation. The fourth level tackles the prevention of unauthorized disclosure of information to an entity actively searching for it using sophisticated means with extended resources, identification and authentication control system specific skills and high motivation.

Group	Assessment Catalogue Question
1. Identification and Authentication Control	
1.1	Do human users have to identify and authenticate themselves on all interfaces?
1.2	Is the segregation for duties and the configuration of least privileges provided?
1.3	Is the utilization of applicable security policies and procedures supported?
1.4	Are human users uniquely identified and authenticated?
1.5	Are all software processes and devices uniquely identified and authenticated?
1.6	Is it possible to configure the strength of a password for password-based authentication(s)? • Is a minimum number of characters required? • Is a predefined variety of different character types required?
1.7	Is the login limited by a configurable number of consecutive invalid access attempts by any user during a configurable time period?
1.8	Is the access of a user denied for a specified period of time or until unlocked by an administrator when this limit has been exceeded?
1.9	Are the minimum and maximum lifetime restrictions of a password enforced for all users (human, software process, or device)?
2. Use Control	
2.1	Are authorization, monitoring and restriction mechanisms for the use of wireless interfaces according to commonly accepted security industry practices supported?
2.2	Are authorized users restricted according to their assigned responsibilities and least privileges?
2.3	Do remote sessions terminate either automatically after a configurable time period of inactivity or manually by the user who initiated the session?
2.4	Is the mobile code verified with integrity/authenticity checks before the code is executed?
2.5	Is it possible to determine whether a given human user took a particular action?
2.6	Is the time source protected against unauthorized alteration and is an audit event caused upon alternation?
3. System Integrity	
3.1	Is the integrity of transmitted information protected against unauthorized manipulation or modification?
3.2	Are cryptographic mechanisms implemented to verify the authenticity of received information and to recognize changes to information during communication?
3.3	Is the data of input fields which directly impact actions validated in accordance to syntax, length and content?

3.4	Are configurable entities automatically notified upon discovering discrepancies during integrity verification?
3.5	Do embedded devices verify the integrity of the firmware, software, and configuration data needed for the component's boot and runtime processes prior to use?
4. Data Confidentiality	
4.1	Is the confidentiality of (persistent and transmitted) information protected for which explicit read authorization is supported?
4.2	Is it possible to delete all information for which explicit read authorization is supported by components that are to be released from active service and / or taken out of service?
4.3	Are volatile shared memory resources protect against unauthorized and unintended information transfer, because data was not purged before the memory was released back?
5. Restricted Data Flow	
5.1	Is the segmentation of networks supported?
5.2	Is the usage of segmented networks supported?
5.3	Is the network traffic denied by default and allowed by exception?
6. Timely Response to Events	
6.1	Are authorized humans and/or tools able to access audit logs on a read-only basis?
7. Resource Availability	
7.1	Is the system able to operate in a degraded mode and maintain essential functions available during a denial of service event?
7.2	Is the load of the communication managed to mitigate the effects of information flooding types of denial of service events?
7.3	Is a backup process supported to safeguard user- and system-level information without affecting normal operations?
7.4	Is it possible to restrict the use of unnecessary functions, ports, protocols and/or services?
7.5	Is the reliability of the backup mechanisms and the integrity of the backed-up information checked before a restore of that information?

Table 1 Assessment Questions dealing with Security

3. Assessment Catalogue for Safety

Table 2 summarizes the safety related assessment questions grouped by the relation to the ISO norms 10218-2 and 13482.

ISO 10218-2:2011[3] "Robots and robotic devices — Safety requirements for industrial robots — Part 2: Robot systems and integration" comprises a specification of safety requirements based on ISO 10218-1 for integrating industrial robots, industrial robot systems and industrial robot cells. All steps ranging from design to maintenance of industrial robots are covered as well as related information and component devices. Basic dangers are described in order to identify requirements for risk reduction, also in the context of integrated manufacturing systems. Process hazards such as ejected chips or welding smoke are not covered in ISO 10218-2:2011.

ISO 13482:2014[4] "Robots and robotic devices — Safety requirements for personal care robots" deals with requirements and guidelines for the usage of personal care robots (mobile servant, physical assistant or person carrier robot) in relation to design safety, protection measures and usage information. In specific, personal care robots are for improving the quality of life of potential users. ISO 13482:2014 focuses on minimizing risks related to such robots. In particular, applications with human-robot physical contact, significant dangers for the types of personal care robots and robot devices within personal care applications are covered. Following aspects are not included in ISO 13482:2014, robots that are not bounded to earth, robotic toys, robots for water or air areas, industrial robots, robots in medical, military, manufacturing or public force context, as well as robots with a speed of more than 20 km/h. The major focus of ISO 13482:2014 lies on human related dangers.

Group	Assessment Catalogue Question
1. ISO 10218-2:2011 Robots and robotic devices — Safety requirements for industrial robots — Part 2: Robot systems and integration	
1.1	How high is the average probability of a dangerous breakdown per hour?
1.2	How high is the medium time until a potential event of a dangerous breakdown?
1.3	Were approved physical parts as well as approved safety measures used for the design and creation of the safety related part of the control system?
1.4	Are the functions of the safety related parts of the control system regularly tested by the machine control?
1.5	Was there a holistic risk assessment conducted to evaluate if there is another safety related performance capability of the control system required for the actual application? (in addition to ISO 10218-2:2011)
1.6	Were the robotic system and the safety measures for robots created with consideration of environmental conditions? (such as temperature, humanity, electro-magnetic failure, light, …)
1.7	Is it possible to manage control and equipment elements from a safe are during the automated operation?
1.8	Are the control and equipment elements located so that visibility to the robot area is ensured?
1.9	Are the control elements aligned with the requirements stated in IEC 60204-1?
1.10	Does the robotic system react on external commands and conditions? • If yes, might this lead to dangerous situations?
1.11	Are the requirements for energy sources for the operation of the robots and additional equipment aligned with the declarations of the producers?
1.12	Is collaborative operation of robots only used for predefined tasks?
1.13	Is collaborative operation of robots possible without activating necessary safety measures?
1.14	Are the collaborative operations aligned with the applications stated in ISO 10218-1?
1.15	Was there a risk assessment conducted for both the collaborative tasks as well as the whole work environment?

[3] https://www.iso.org/standard/41571.html (last visited on 18/11/2020)

[4] https://www.iso.org/standard/53820.html (last visited on 18/11/2020)

1.16	Are the robots that are integrated in a collaboration environment aligned with the requirements of ISO 10218-1?
1.17	Is there an integrated safety feature for the detection of presence? • If yes, are the requirements fulfilled?
1.18	Are there integrated safety features? • If yes, are the requirements fulfilled?
1.19	Can the technical safety measures prevent or recognize the continuous approaching of a person within the safe area?
1.20	Can the technical safety measures stop the robots and dangerous functionalities immediately in case of a person is approaching the safe area?
1.21	Can external safety measures prevent or recognize the access of people in areas that are not designated for collaborative operation?
1.22	Are there other machines, connected to the robotic system, that might lead to a potential danger in the same collaboration area? • If yes, are those machines aligned with the required safety functionalities?
1.23	Is there a clear marking of collaboration areas, where humans interact with robots? (such as warning lines or signs)
1.24	Is operating personnel safe by a combination of integrated safety functionalities and the compliance with robotic parameters, as well as to dangerous actions stop in case of emergency?
1.25	Is there more than one person (operating personnel) involved in the collaborative operation? • If yes, is the safety of all people ensured by individual control elements?
1.26	Is the collaboration area designed in a way that operating personnel can conduct all tasks easily and that the positioning of equipment and machines does not increase the danger?
1.27	Are there additional safety measures available that prevent bruising and catching in areas with less than 500 mm open space?
1.28	Are there dangerous situation (for people) when changing the operating mode from autonomous to collaborative?
1.29	Is there a safety stop after each recognized breakdown of safety characteristics for collaborative operation?
1.30	Is it possible to restart the autonomous operation by scienter after a safety stop? • If yes, can such a restart only be triggered from outside of the collaboration area?
1.31	Does an autonomous robot operating in a collaborative area stop when a person enters the area in order to enable direct interaction between operating personnel and robot?
1.32	Is there a safety stop during the hand-operated operation, when the robot reaches the position for handover?
1.33	Has the operating personnel a guiding device for hand-operated operation to transfer the robot to a specific position? • If yes, are the requirements fulfilled?
1.34	Has the operating personnel a clear sight on the whole collaboration area during the hand-operated operation?
1.35	Is there a safety related stop in case the operating personnel drops the guiding device?
1.36	Are only robots aligned with ISO 10218-1 used in a system that is designed in a way that a safe distance between operating personnel and robots can be ensured in a dynamic operation process?
1.37	Are the speed of the robots, the minimal distance and other parameters defined in the risk assessment?
1.38	Are only robots aligned with ISO 10218-1 used for systems that are designed to control dangers by limitation of energy and strength/power?
1.39	Are parameters for performance, strengths/power and ergonomics defined in the risk assessment?
2. ISO 13482:2014 Robots and robotic devices — Safety requirements for personal care robots	
2.1	Is the charging constructed in a way that by chance touching of current-carrying parts is prevented?
2.2	Does the usage description describe the risks related to breakdown or shut down of the energy supply?
2.3	Is ensured that the robot checks during the start-up if all safety related functionalities are available?
2.4	Is there a safety stop if safety related functionalities cannot be tested during the start-up?

2.5	Are manipulators and other mobile elements deactivated during the start-up?
2.6	Which safety measures were applied? • software controlled limitation of the assistant robot's working area • limitation of speed and safety related monitoring of speed • limitation of strengths/power and safety related monitoring of strength/power
2.7	Does the risk assessment indicate that a missing perception of the robot by the human can be a danger?
2.8	Which of the following safety signals are applied? • acoustic signals to warn the user against potentially dangerous situations • light signals or other optic features to warn users and third-party actors against the presence of personal assistant robots • safety stops while a relevant object is within the safety area of a robot, the robot proceeds with the tasks after the object left the safety area

Table 2 Assessment Questions dealing with Safety

4. Assessment Catalogue for Legal Aspects

4.1 The role of law, especially criminal law, for trustworthy AI in robotic platforms

For better interdisciplinary comprehensibility, some basic aspects shall be clarified before setting up an assessment catalogue for legal aspects. Within the evolving framework of technical abilities of AI-based systems legal aspects in connection with ethics increasingly play an important role already at the development stage. Legal aspects are important especially for two reasons:

1. For developing a trustworthy AI for the user
2. For legal compliance reasons, especially product safety, and in order to protect the developer in our project from criminal liability, if the AI should cause damage to a person / property, or the AI-based system gives hacking opportunities during the application

To make sure that an AI that is legally compliant, especially criminal law must be taken into account already during the development and production. Persons as well as companies that are involved in the process must act within the frame provided by legal norms in order to avoid damage during the application and, thereby, prevent potential penalties against themselves. Furthermore, by considering legal compliance already at the early stage of developing procedure, it should be ensured that the AI-based system itself is legally compliant. Thus, trust will be established and supported.

The criminal law is predestined as an example of the inclusion of legal framework conditions for the following reasons:

- ***Criminal law compliance establishes trust****:* AI-based systems' compliance to prohibitive criminal law norms helps to build a trustworthy AI-based system. Criminal law protects (or is intended to protect) by means of potential punishment. It aims at complying with due diligence standards. Thereby, it helps to avoid accidents and damages that can arise through the applications of AI-based systems. Accidents might cause potential user discomfort and thereby disrupt the provision of trust. For example, core criminal laws aim at protecting life and limb, physical integrity, property damage, data abuse and hacking.

- ***Criminal law compliance aims at avoiding legal uncertainties for developers and manufacturers****:* Criminal law is also the strictest means of sanctioning actions non-compliant to legal standards and norms. Especially in innovative areas, in early stages of development typically there will not exist more detailed legal norms (or even – because of the rapid technical development possibilities – will never be available) and thus legal uncertainties exist. But developers should be protected from legal uncertainties and risks as for as possible. It is therefore particularly important to act with due diligence and predictable in order to prevent subsequent sanctions.

  It should only be noted here briefly that the domestic laws, especially criminal laws apply typically only within the state's territory. Nevertheless, core domestic rules cover also extraterritorial acting or damages, and comparable rules are provided for in most states (such as the criminal liability for intentional or negligent killing, or the liability for bodily injuries, or other intrusions to legally protected interests as mentioned above). In addition, it can be said that especially in the German-speaking states, comparable approaches such as due diligence and predictability apply in avoiding negligent criminal culpableness. In principle, it can also be said that comparable results are achieved in other countries regardless to differences in dogmatic approaches.

  Certain terms (such as "responsibility", "risk", "danger") are used here from a criminal law perspective, but have also been discussed in advance in an interdisciplinary dialogue in order to clarify the situation within the project team.

4.2 Requirements

A focus on criminal law has been set. As already mentioned, the focus is specifically set on avoiding negligent actions, as the processors assume that intentional action by developers and companies (e.g. believing that damage is serious possible and accepting it) typically does not occur (except when it comes to so-called "dilemma situations", described e.g. by Philippa Foot in the so-called "trolley problem").

Priority is therefore given to the "occurrence" of errors and inaccuracies during the development that are causal for damage in practice.

a) **Focus on developers (producers) and companies**: From a criminal law perspective, several alleged perpetrators can be considered in parallel, i.a. as the developer(s) and the manufacturer, an admission board or the user. Since this is an exploratory project, in our case the developer / manufacturer as well as the corporation should be protected by a criteria catalogue. Hence, only assessment criteria referring to the following groups are developed:

- developer and manufacturer (integrators are included here) as individuals,
- as well the corporation behind it (being a legal entity), as there is already a corporate criminal liability enacted in many countries, especially in Austria. Corporate criminal liability can lead i.a.to large fines for companies, and might be accompanied by other legal orders and consequences.

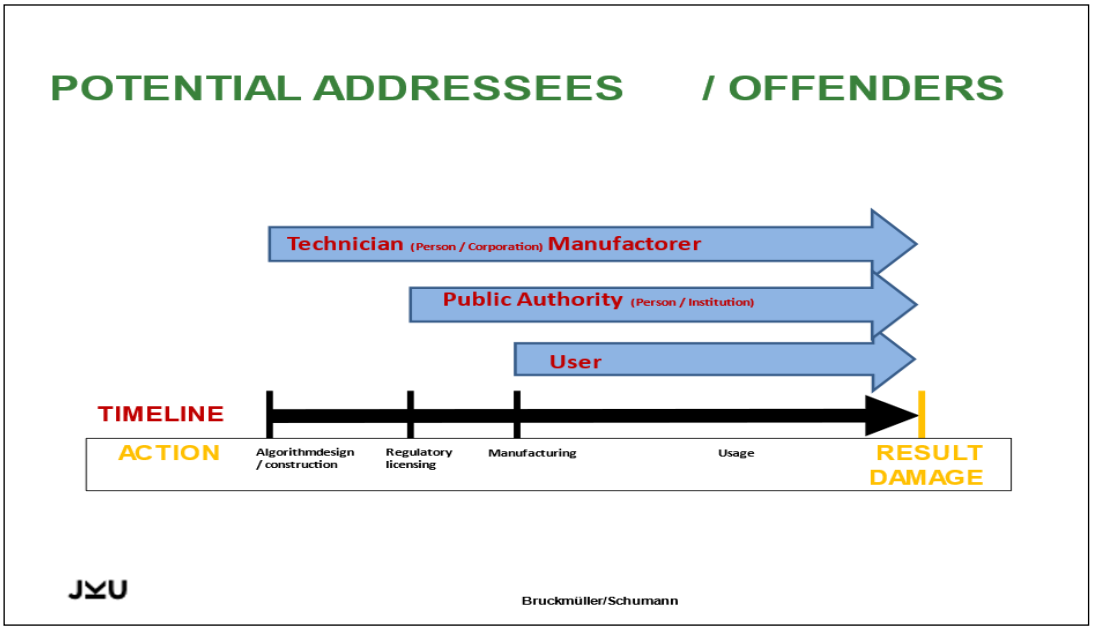

Figure 1 Sample of Potential Addresses of Criminal Liability in case of AI-based Systems

b) **Prevention of criminal liability for negligence** and questions formulated as a criteria catalogue:

- prevention of non-due diligence and
- prevention of risks (predictability)

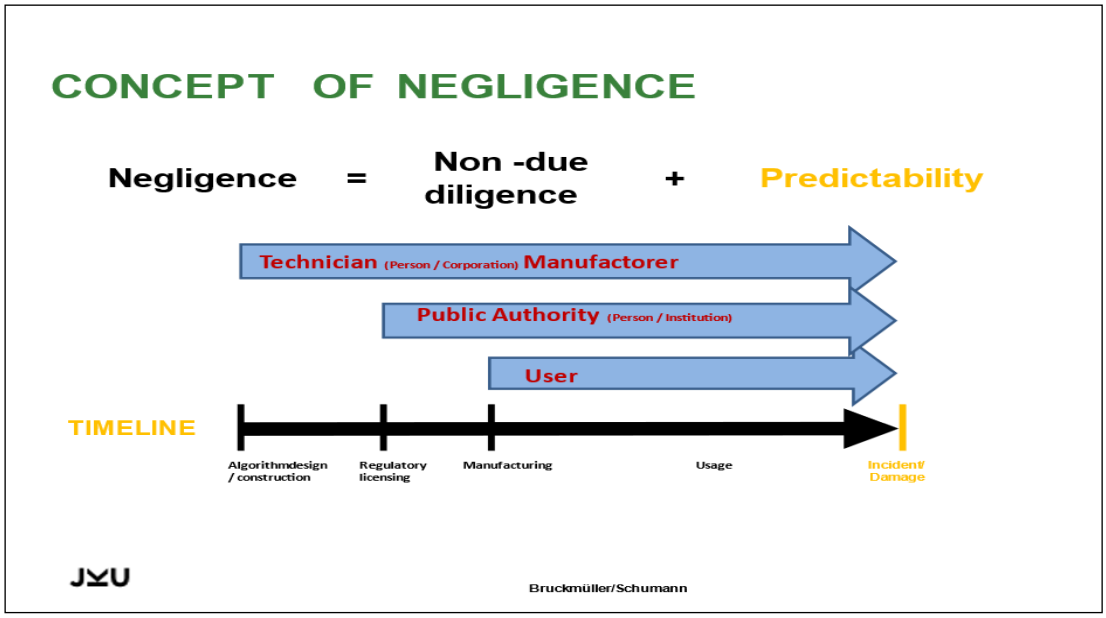

Figure 2 The Concept of Negligence: Non-Due Diligence + Predictability

- The required standard of due diligence needs to be established from an objective perspective. Individual disabilities might only limit individual criminal liability but do not limit the objective standards to be met.

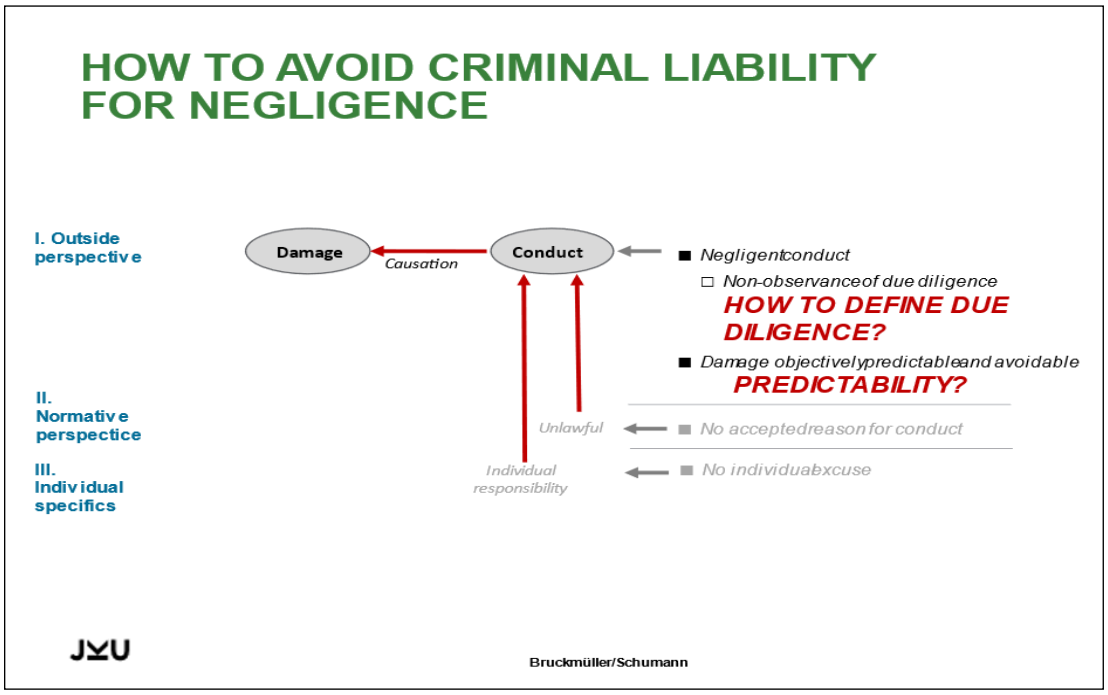

Figure 3 Needs in order to avoid Criminal Liability for Negligence

- There is a graduated system of establishing the necessary standard:

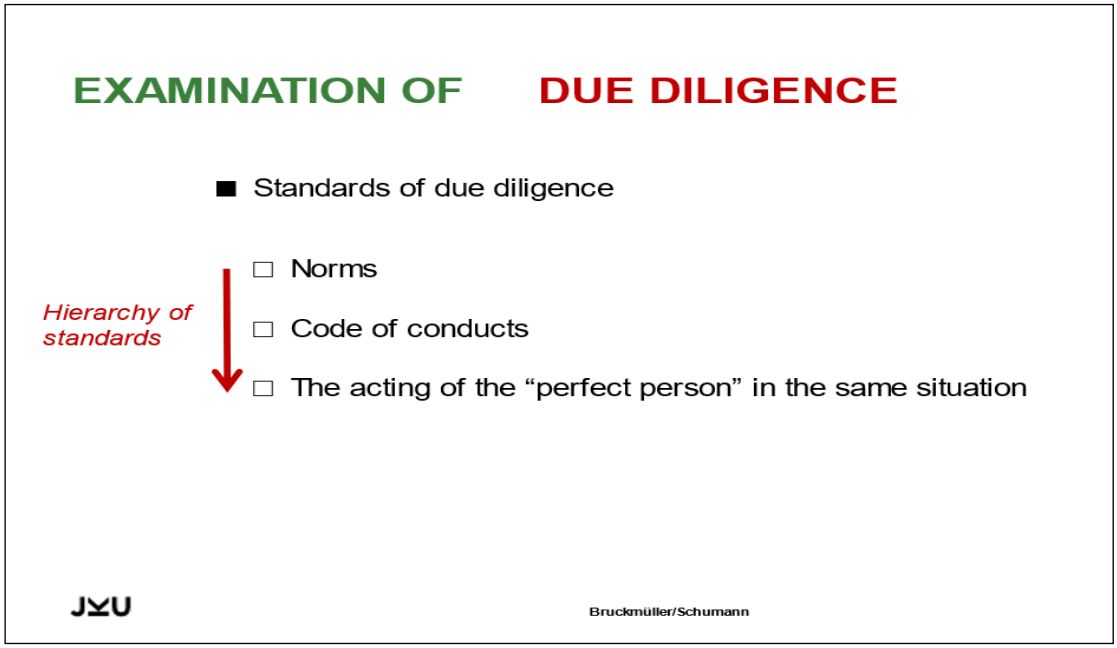

Figure 4 Basic Legal Definition of Due Diligence

- The predictability of potential risks, in general, limits potential accusations of negligent acting. However, especially the implementation of innovative methods and systems demands for supervision of these products. Supervision enables to detect potential risks in process. Thus, supervision forms part of due diligence standards.

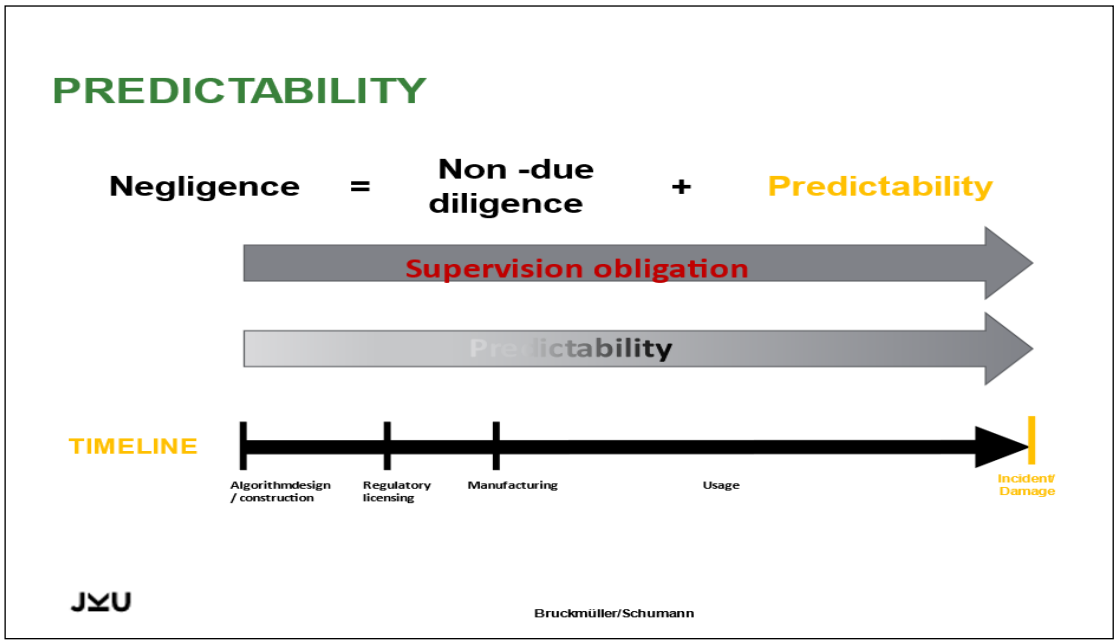

Figure 5 Interrelation of Predictability and Supervision Obligations

- IMPORTANT NOTE: Full exclusion of risks of criminal liability cannot be guaranteed; neither by this study which it is primarily a technical study and is only started selectively in order to test technical possibilities for the implementation of the technical, legal and ethical criteria catalogues, nor in general. Furthermore, data protection issues were not taken into account in this study.

c) **Questions based on use cases**: Criminal law is based on (potential) concrete harmful results, action, or damage. The questions asked here on the basis of the developed use cases aim to avoid such criminal results. Where different opinions or methodologies exist, questions aim to comply with the requirements established by Austrian jurisprudence and prevailing opinion.

The underlying use case (for BOC and Johanneum Robotics): The robot arm is set up in a secured room / area with the corresponding goods in its environment. In principle, no person is allowed to enter the room / area when the robot is "working". Nevertheless, an employee walks into the room while the arm is holding goods, although this is prohibited. The robot arm hurts the human. However, the robot arm by itself must act in such a way that the person is protected and not injured, if necessary, the arm must stop or move in another direction, even if objects may be destroyed as a result. The assessment catalogue question will focus on this use case since this case in particular covers the relevant legal aspects.

The project partners have agreed that the ethically possible corridor will be limited by criminal law, so in the event of contradictions or areas of conflict between law and ethics, criminal law will take precedence - in order to prevent sanctions.

Table 3 summarizes the assessment questions related to the above explained legal issues grouped by key aspects including case specific questions (indented).

Group	Assessment Catalogue Question
1. Objective acting with due diligence	
1.1	Is the acting during the development objectively in line with due diligence standards?
1.1.1	Is the acting in line with all relevant *legal norms*? (e.g. avoidance of *Negligent bodily injury*, § 88 Austrian Criminal Code "Fahrlässige Körperverletzung") • Case-specific: Are sufficient measures taken to ensure that no person (also when s/he enters unauthorized) will be bodily injured?
1.1.2	Are all the *standards / leges artis* followed? • Case-specific: Have all the relevant safety norms been followed in order to avoid that persons are bodily injured? • Especially, have ISO 10218-2:2011 and ISO 13482:2014 been followed?
1.1.3	Would the "perfect person" (meaning a thoughtful and prudent person of the same professional group or role) act in the same way when developing an AI system like this? • Case-specific: Have all necessary safety measures been taken which deem to be necessary to avoid injuries?
1.2	Is it necessary to overrule standards in order to ensure that innovative creations are safe? • Case-specific: Is it (really) necessary to overrule a standard to prevent that a person is bodily injured?
1.3	Have the relevant steps been documented for evidence purposes? • Case-specific: Was the reasoning/justification for or against the procedure decisions documented (pros and cons)?

2. Predictable risks assessment: Assessment of existing risks in the application; Conduction of risk assessment; Consideration of future dangerous situations	
2.1	Have all the foreseeable / predictable risks emanating from the AI system in action been assessed in the development process? • Case-specific: Have all the potential risks been analysed that could result in bodily injury of a person using the arm?
2.2	Have all other dangerous situations been assessed which might occur due to the usage of the AI system? • Case-specific: Have any possibly dangerous situations been considered that could arise from the *user / buyer* of the arm? • Have any possible dangers that could arise from *third parties* who come too close to the robot arm been considered? • Have any possible dangerous situations been considered that could arise from possible *future locations* of the robot arm? • Have the risks / dangerous situation been balanced against each other and were decisions made in such a way that people are not injured (is priority given to protect humans before goods)? • Have all the evaluation decisions and reasonings been documented?
3. Technological measures preventing the realization of risks / dangers	
3.1	Have all technical countermeasures been taken to avoid risks and thus prevent damage? • Case-specific: Have all technical possibilities been exhausted so that no risk arises? • Have all technical possibilities been exhausted to prevent dangers from the AI system environment? • Have mechanisms been built in so that the robot arm stops immediately when a person comes close (e.g. safety contacts that stop the robot as soon as the door / fence is opened)? • Has an off-button been installed? Is it possible for a human to take over the robot arm control in case of emergency? Here, too, check if the corresponding safety standards have been installed.
4. Information of the (future) user in order to prevent realization of risks / dangers	
4.1	Has the integrator or user been informed about measures to prevent possible damage? • Case-specific: Has the integrator been informed about necessary security measures in the application environment? • Has an integrator been selected who can assess the risks posed by the environment? • Has the user / buyer been informed about risks and how to avoid them? • Will the arm be updated when the environment changes?
5. Correct balancing of legal interests (outside of dilemma situations)	
5.1	Is it ensured that in a case of an emergency the higher legal value is always protected before or instead of the minor value? • Case-specific: Was the arm designed to always protect humans before goods/things? • Thus, generally spoken, maybe destroy a thing in an emergency case, but do not injure a person!
6. Subjective due diligence	
6.1	Has care been taken to ensure that the persons involved are mentally and physically capable of objectively acting with due diligence? No negligence of taking over potentially dangerous tasks if one cannot follow the necessary standards? • Case-specific: Has the development been entrusted to such persons who are also appropriately trained? (In addition, see the questions to the corporation/legal entity below.)
7. Selection and supervision of personnel	
7.1	Given the requirements defined before, are the personnel involved in developing and applying the AI based system selected in line with these requirements? • Case specific: Is the personnel involved sufficiently qualified and capable of fulfilling its tasks?
7.2	Are the personnel involved sufficiently informed (especially about its responsibilities and, as far as necessary, about the distribution of tasks) and trained?
7.3	Are the personnel involved sufficiently supervised?

7.4	Is the establishment of a whistle-blower system (or similar) outside of the reporting line evaluated in order to ensure that difficulties and potentially risky developments can be easily and anonymously be reported in house?

Table 3 Assessment Questions dealing with Legal Aspects

5. Assessment Catalogue for Ethics

Transdisciplinary integration was fostered based on synchronization of pre-knowledge as well as understanding of the problem and translation into a cross-disciplinary (ethical) research question - by means of team internal expert interviews. Using personal expectations and fears on technical futures as "boundary objects" in order to generate a mutual normative understanding. Where a "boundary object" is an intuitively given methodical vehicle that functions as some kind of integrative focal point (Bergmann et al. 2010, pp. 106-116[5]; Klein 2017; Pohl et al. 2008[6], pp. 415-416; Mittelstraß 2018[7], p. 70). Therefore, the following team internal requirements/values have been [8]identified:

- Quality of technical products is more important than its quantitative propagation
- Respect end user´s needs and don't abuse them as human guinea pigs
- Social wellbeing prior to economic benefit
- Human-cantered AI-development
- Overcoming negative impacts of both factual constraints ("Sachzwänge") and impossible backward compatibility ("ausbleibende Rückwärtskompatibilität") so that technological issues do not limit or prefigure human decision making – also in future generations
- Risk Management as complex task in technical systems including Robotics and AI

In a constructive sense some basic terms became the objects of critical controversies within the team. Although no final definition has been found, the debates based on "boundary objects" served as integrative method in order to facilitate mutual problem orientation within the team. Throughout the first meetings, the ethics and law teams tried to converge on some of the more complex ideas and concepts underlying these critical controversies. This somewhat reflects the 'reflective equilibrium'-approach (Daniels 2020)[9], in which both teams tried to find common denominators to address these terms. For this, they also relied on existing approaches from (Bio)Medical Ethics (Beauchamp & Childress 2001)[10], which influence current attempts to regulate the ethical implementation of AI (AI HLEG 2019). The terms "risk", "danger", "responsibility", "robot", "ethical corridor", and "AI" were found to be particularly ambiguous and contentious. Some of the most important findings can be summarised as follows:

- In ethics, there is a tendency towards approaching responsibility as a wide and flexible concept (i.e. depending on different factors and principles). In law, however, this is not the case. Here, responsibility has to be understood as narrow and instrumental concept (i.e. as a means to an end). It was hence particularly

[5] Bergmann, Matthias, Thomas Jahn, Tobias Knobloch, Wolfgang Krohn, Christian Pohl, Engelbert Schramm 2010: Methoden transdisziplinärer Forschung. Ein Überblick mit Anwendungsbeispielen. Frankfurt a.M. & New York: Campus.

[6] Pohl, Christian, Lorrae van Kerkhoff, Gertrude Hirsch Hadorn & Gabriele Bammer 2008: "Integration." In: Hirsch Hadorn, Gertrude, Holger Hoffmann-Riem, Susette Biber-Klemm, Walter Grossenbacher-Mansuy, Dominique Joye, Christian Pohl, Urs Wiesmann, Elisabeth Zemp (eds.) 2008: Handbook of Transdisciplinary Research. Dordrecht: Springer, pp. 411-424.

[7] Mittelstraß, Jürgen 2018: "The Order of Knowledge: From Disciplinarity to Transdisciplinarity and Back." In: European Review, Vol. 26, No. S2, pp. 68–75. [doi:10.1017/S1062798718000273]

[8] Klein Julie Thompson 2017: "Typologies of Interdisciplinarity. The Boundary Work of Definition." In: Frodeman, Robert, Klein, Roberto C.S. Pacheco (eds). 2017: The Oxford Handbook of Interdisciplinarity. Second Edition. Oxford: Oxford University Press, pp. 21-34.

[9] Daniels, N. (2020). Reflective Equilibrium. The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/.

[10] Beauchamp TL, Childress JF (2001) Principles of Biomedical Ethics. Fifth Edition. Oxford University Press, Oxford

difficult to find a common understanding for “responsibility”, which can be used from both sides, ethics and law.

- There seemed to be quite some ambiguity concerning the understanding, the use and the relation of the terms “risk” and “danger”[11]: team law differentiated between *direct danger*, which is assessable, and *abstract danger*, which is recognisable, yet not necessarily known). Risk, however, was understood as very difficult to assess. However, while this is one view that was presented in both the expert interviews and subsequent discussions, there are various other views on this, e.g. risk as quantifiable (risk assessment), and danger as rationalizable. From the ethics side, both risk and danger were understood to be more flexible, especially with regards to the methodology in which both can be addressed. Again, it rendered quite difficult to find a common understanding of how these two terms relate to one another, especially because of their flexibility (even within the respective disciplines).

Lessons learned: from the technical and model-based side, the project can indeed succeed. However, there were some challenges from the ethical and juridical side, since the criteria, which are to be implemented in the criteria catalogue remain rather abstract. This is largely due to the many controversies found throughout the ‘reflective equilibrium’-approach. It is based on this, that in order to specify ethical criteria, one needs to look at specific use cases, and evaluate the ethical implications based on these.

With this, one takeaway for future projects is that it remains highly challenging to bring together ethical criteria with law. To find a most efficient and reliable way to develop an ethics and law criteria catalogue, it is hence advisable to elaborate use cases, and define the questions for the respective criteria catalogue individually, for each implementation scenario.

Following settings were assumed for the specific case. The robot arm is installed in a (secured) room/area with the relevant items placed in its environment. In principle, human agents are not allowed to enter the room/area when the robot is operating. However, although this is prohibited, an employee enters the room while the robot arm is carrying goods. The robot arm causes an injury to the person. In principle, it must act in such a way that it protects the human agent (who in this case is an employee) and does not injure her, stops if necessary or moves into a different direction, even if thereby destroying other items/objects.

The ethical assessment is limited to the perspective of design and production. Maintenance, distribution, marketing, implementation and application are blended out. Therefore, the main ethical question is: What are concrete criteria/requirements that need to be taken into account within an AI system in collaborative robotics?

Following Sources serve as a foundation for developing the ethical criteria catalogue.

- ACRAI (2018)[12] Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positiv gestalten. White Paper des Österreichischen Rats für Robotik und Künstliche Intelligenz. Wien, November 2018
- AI HLEG (2019)[13] Ethics Guidelines for Trustworthy Artificial Intelligence. High-Level Expert Group on Artificial Intelligence. 8. April 2019. European Commission, Brussels

[11] The findings of risk vs. danger are the result of the conducted expert interviews. It is hence important to emphasise that these aspects, first and foremost, represent intuitive answers concerning possible ambiguities and challenges.

[12] Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positiv gestalten. White Paper des Österreichischen Rats für Robotik und Künstliche Intelligenz. Wien. Retrieved from https://www.acrai.at/wp-content/uploads/2020/03/ACRAI_White_Paper_DE.pdf (last visited on 21/12/2020)

[13] AI HLEG (2019). Ethics Guidelines for Trustworthy Artificial Intelligence. High-Level Expert Group on Artificial Intelligence. European Commission, Brussels. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. (last visited on 21/12/2020)

- AI HLEG (2020)[14] The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. 17. July 2020. European Commission, Brussels

- IEEE (2019)[15] Ethically Aligned Design. Online Document.

Table 4 summarizes the ethics related assessment questions – for the design and development of AI systems – grouped by key aspects including case specific questions (indented). The criteria are provided in order to be tested and applied within a concrete technical case study by BOC and JR that is synchronized with legal issues (JKU Linz).

Group	Assessment Catalogue Question
1. Human Agency & Autonomy (AI HLEG 2019 & 2020)	
1.1	Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision? Case-specific: Is there any form of human-AI *interaction* in place that allows for mechanism in case the system places the wrong product in the shopping cart, and another product needs to be put back?
1.2	Could the AI system affect human autonomy by interfering with the end user's decision-making process in any other unintended and undesirable way? Case-specific: Does the system stick to the shopping list of the human shopper? Or can it make suggestions according to the shop's incentives?
2. Human Oversight (AI HLEG 2019 & 2020; IEEE 2019)	
2.1	Is there a "human in control"? If yes, what level of control: *Human*-in-the-Loop or *Human*-on-the-Loop? Case-specific: Is there a human agent overseeing the assembling process of the robot arm?
2.2	Is there a procedure to safely abort an operation where needed? Case-specific: If a human agent enters the room while the arm is operating, and keeps operating (malfunction), does she have the possibility of stopping the robotics arm, even though she is injured, from any given place in the room?
2.3	Does this procedure abort the process entirely, in part, or delegate control to a human? Case-specific: Who or what can stop the robot arm from assembling items?
3. Technical Robustness & Safety (AI HLEG 2019 & 2020; IEEE 2019)	
3.1	Does the system fully comply with the criteria given in the safety & security catalogue? Case-specific: (See technical safety standards applied to the system.)
4. Privacy & Data Governance (AI HLEG 2019 & 2020; IEEE 2019)	
4.1	Are there mechanisms in place that ensure the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection? Case-specific: Is the data protection of the human shopper ensured? (personal data and data on shopping preferences)
4.2	Is the system being trained, or was it developed, by using or processing personal data (including special categories of personal data)? Case-specific: Is the system being trained, or was it developed, by using or processing personal data (including special categories of personal data)? What kind of shoppers and what circumstances does this initial training data represent (e.g. different shopping behaviour during COVID)? Does the system fully comply with the criteria given in the safety & security catalogue?
4.3	Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?

[14] AI HLEG (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. European Commission, Brussels. Retrieved from https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-assessment-list-trustworthy-artificial-intelligence-altai. (last visited on 21/12/2020)

[15] IEEE (2019). Ethically Aligned Design. Online Document. Retrieved from https://standards.ieee.org/industry-connections/ec/ead-v1.html (last visited on 21/12/2020)

5. Transparency (AI HLEG 2019 & 2020; IEEE 2019)	
5.1	Are there mechanisms that allow the developers to trace back which data was used by the AI system to make a certain decision? Case-specific: Can the developer trace back why the robot chose to give human shopper item x, instead of item y?
5.2	Are there measures to continuously assess the quality of the output(s) of the AI system? Case-specific: Is there a trained human agent, who continuously assesses whether robot arm outputs put human shopper's good first?
5.3	Do developers continuously survey the users if they understand the decision(s) of the AI system? Case-specific: Do developers survey human shoppers if they understand why robot arm chose items x,y,z?
5.4	Are there mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system? Case-specific: Does the supermarket communicate the benefits of the AI system to users? Does the supermarket communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?
6. Diversity, Non-Discrimination & Fairness (ACRAI 2018; AI HLEG 2019 & 2020; IEEE 2019)	
6.1	Does the system foster racial bias? Case-specific: Does the majority of human agents, who performed item assembling processes before atomizing them, identify with an underrepresented ethnicity?
6.2	Does the system foster gender bias? Case-specific: Does the majority of human agents, who performed item-assembling-processes before atomizing them, identify with an underrepresented gender?
6.3	Does the system foster socio-economic bias? Case-specific: Does the majority of human agents, who performed item assembling processes before atomizing them, identify with a specific socio-economic background?
6.4	Are there educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system? Case-specific: Are there educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system?
6.5	Is the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion? Case-specific: Does the ordering process exclude certain groups of society? (= is the use of the services provided by the robot arm accessible to everyone?)
7. Societal & Environmental Well-Being (AI HLEG 2019 & 2020; IEEE 2019)	
7.1	Are there potential negative impacts of the AI system on the environment? Case-specific: If the system is not used, does it go into 'energy-saving' mode?
7.2	Are there measures to reduce the environmental impact of the AI system throughout its lifecycle? Case-specific: Does the robot arm run on 'green' electricity? Can parts of the robot arm be recycled?
7.3	Does the AI system impact human work and work arrangements? Case-specific: Does the robot arm replace human workers? (see also point 6) Diversity, non-discrimination and fairness) Does the system promote or require new (digital) skills?
8. Accountability (AI HLEG 2019 & 2020; IEEE 2019)	
8.1	Can the AI system be audited by independent third parties? Case-specific: Is there a third party, which regularly audits the whole system?
8.2	Is there risk training for the human agents involved, and, if so, does this also inform about the potential legal framework applicable to the AI system? Case-specific: Is there risk training for the human agents involved, and, if so, does this also inform about the potential legal framework applicable to the AI system? What measures are taken to ensure that no one enters the room when the robot arm is operating? (e.g. warning signs, sirens, etc.)

9. Responsibility (Secondary Importance for the Case Study)	
9.1	Is the attribution of responsibility to concrete human agents possible? Personally and socially embedded attribution of responsibility both on the individual and collective level ACRAI 2018).
10. Values (Secondary Importance for the Case Study)	
10.1	Human Rights, Human Dignity, Privacy, Democracy and Participation (ACRAI 2018; IEEE 2019) etc.

Table 4 Assessment Questions dealing with Ethics

6. Conclusion

This document at hand summarizes the assessment catalogue questions created in collaboration with robotic engineers, law and ethic experts. The content belongs to work package 5 “Erstellung eines Kriterienkataloges für Ethische-, Rechtliche und Sicherheitsfragen”.

Summarizing, the questions are classified in four criteria catalogue types. Those are security, safety, legal and ethics questions. The catalogue types contain questions that are further subdivided in key aspects or ISO norms for instance. The legal as well as ethics questions are divided in general questions which are further specified in case specific context.