Assessing Trustworthy Artificial Intelligence of Voice-enabled Intelligent Assistants for the Operator 5.0

Alexandros Bousdekis[1], Gregoris Mentzas[1], Dimitris Apostolou[1,2], and Stefan Wellsandt[3]

[1] Information Management Unit (IMU), Institute of Communication and Computer Systems (ICCS), National Technical University of Athens (NTUA), Athens, Greece
{albous, gmentzas}@mail.ntua.gr
[2] Department of Informatics, University of Piraeus, Piraeus, Greece
dapost@unipi.gr
[3] BIBA - Bremer Institut für Produktion und Logistik GmbH at the University of Bremen, Bremen, Germany
wel@biba.uni-bremen.de

Abstract. The concept of Trustworthy Artificial Intelligence (TAI) focuses on the establishment of trust in AI systems' development, deployment, and use. In this realm, the European Commission (EC) developed the Assessment List for Trustworthy Artificial Intelligence (ALTAI) in order to enable the assessment of trustworthiness in the AI systems under development. Since this is an emerging topic, there is little evidence on how to apply ALTAI. In this paper, we present the application of ALTAI on a Digital Intelligent Assistant (DIA) for manufacturing. In this way, we aim at contributing to the enrichment of ALTAI applications and to the drawing of remarks regarding its applicability to diverse domains. We also discuss our responses to the ALTAI questionnaire, and present the score and the recommendations derived from the ALTAI web application.

Keywords: Trustworthy AI, voice assistant, AI ethics, Assessment List for Trustworthy Artificial Intelligence, ALTAI, Industry 5.0.

1 Introduction

To maximize the benefits of Artificial Intelligence (AI), while at the same time mitigating its risks and dangers, the concept of Trustworthy AI (TAI) promotes the idea that individuals, organizations, and societies will be able to achieve the full potential of AI if trust is established in its development, deployment, and use [1]. The TAI concept has been studied in several works [2-6]. The increasing literature implies that ethics have been put at the core of the development of AI technologies [7], especially of those that incorporate predictive capabilities [4,8]. In European Commission (EC)'s strategy, published in 2018, AI must be lawful, ethical, and robust [1,9], and is defined as "systems that display intelligent behaviour by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals". In this context, EC developed the Assessment List for Trustworthy Artificial Intelligence (ALTAI), "a

practical tool that helps business and organizations to self-assess the trustworthiness of their AI systems under development" [10]. Since this is an emerging topic, up to now there is little evidence on how to apply ALTAI [11].

In this paper, we present the application of ALTAI to a Digital Intelligent Assistant (DIA) for manufacturing. DIAs represent a new type of interaction between operators and machines in the context of Industry 5.0 aiming at establishing mutually beneficial relationships between smart technologies, and the Operator 5.0 [12]. The DIA was developed in the COALA ("COgnitive Assisted agile manufacturing for a LAbor force supported by trustworthy Artificial Intelligence") EU research project which provides a proactive and pragmatic approach to support operative situations characterized by high cognitive load, time pressure, and zero tolerance for quality issues. For more details on the COALA concepts and technologies, the reader may refer to [13-15].

The rest of the paper is organized as follows: Section 2 briefly reviews the main frameworks for TAI. Section 3 reviews the related works on applications of ALTAI. Section 4 discusses our responses to the ALTAI questionnaire, and presents the score and the recommendations derived from the ALTAI web application. Section 5 presents the main concluding remarks on the applicability of ALTAI, while Section 6 concludes the paper.

2 Review of Frameworks for Trustworthy AI

During the last years, several frameworks have arisen referred to as Beneficial AI [16], Responsible AI [17,18], Ethical AI [2,19], and Trustworthy AI [1,20]. An overview of some of the most representative works is presented in **Table 1**. Despite their value for TAI realization, they exhibit two main limitations [6]: (i) Several TAI principles may conflict with each other, depending on the application cases; (ii) They are general and they do not provide sufficient guidance on how they are transferred into practice.

Table 1. Description of the main TAI frameworks.

Framework	Description
Asilomar AI Principles [16]	23 principles of beneficial AI, organized into three categories: research issues, ethics and values, and long-term issues.
Montreal Declaration of Responsible AI [17]	10 ethical principles that promote the fundamental interests of people and groups, and 8 recommendations for responsible AI.
UK AI Code [19]	5 principles for an ethical AI code, intended to position the UK as a future leader in AI.
AI4People [2]	Synthesis of 6 frameworks, which resulted in 5 foundational principles for ethical AI, and a set of 20 action points.
OECD Principles [20]	5 principles for the responsible stewardship of trustworthy AI.
Governance Principles for the New Generation AI [18]	A framework and action guidelines for the governance of AI, based on 8 principles for the development of responsible AI.
EU Ethics Guidelines for TAI [1]	4 principles and 7 key requirements for achieving TAI. An assessment list for the operationalization of the requirements.

3 Related Works on Applications of the ALTAI Tool

Although the Ethics Guidelines for TAI [1] have received some criticism [21,22], they emerge to be influential in EU, since they try to achieve an inclusive consensus of how societies can deal with the opportunities and challenges of AI. They serve as the basis for the regulation of AI, the EU AI Act [23], and the principles for TAI in EU [24]. Whilst the ALTAI list is not the only example of an AI impact assessment [25-27], its visibility benefits from the central role it plays in EU AI policy [28].

Table 2. Remarks from ALTAI applications.

Ref	Main Remarks
[11]	ALTAI variants should be developed for the various software lifecycle phases. Domain-specific adaptations of ALTAI should evolve. ALTAI should be reorganized to support readability.
[29]	ALTAI should entail definitions of widely used terms. ALTAI should have specific versions for different domains. Overlaps and redundancies were found within/throughout the 7 requirements. ALTAI should consider what questions are relevant to which development stage.
[30]	The weighing of each question to the final scores is not clear to the user. The ALTAI Polar diagram presents the results in an unappealing way for early-stage organizations who might score poorly to present their results publicly. ALTAI does not consider the appropriate level of governance. There is no consideration to the relative risk of AI systems in the assessment process. The consequence of the prerequisite Fundamental Rights Impact Assessment (FRIA) failing would negatively affect the ALTAI score. Some consideration should be given to the merits and demerits of the nature of the tool. Organizations need to understand how they compare to their peers as well.
[31]	ALTAI do not publish how each answer contributes to the final scores. The ALTAI recommendations seem extensive, repeated, and difficult to understand. The ALTAI score is difficult to be interpreted in terms of guidelines for improvement.
[32]	ALTAI lacks a clear strategy toward various entities that can be affected by unintended harmful consequences of the AI systems. Relevant ethical issues can be mapped out on how data moves across the AI systems. The ALTAI risk-based approach can be reinforced to categorise the ethical risks and countermeasures in relation to the specific stakeholders.
[28]	ALTAI considers AI as an ethical issue instead of a technology, or family of techniques. ALTAI creates questions of applicability that are independent of the actual ethical and social consequences of the specific AI system under examination. The focus on trustworthiness does not fully represent all the ethical and social concerns. Limitations exist when applying an ex-ante instrument (i.e. ALTAI) at the research stage.

Applications of ALTAI include: Advanced Driver-Assistance System [11], Early Warning System in Education [4], AI-supported Air Traffic Controller Operations [29],

AI-based technologies for ageing and healthcare [30], neuroinformatics [28]. **Table 2** presents the main derived remarks from such applications.

4 Application of ALTAI in Digital Intelligent Assistants for Manufacturing

4.1 Analysis and Discussion

The Ethics Guidelines for TAI [1], published by the High-Level Expert Group on Artificial Intelligence (AI HLEG), contains an Assessment List to help assess whether the AI system adheres to the seven requirements of TAI: (1) Human agency and oversight; (2) Technical robustness and safety; (3) Privacy and data governance; (4) Transparency; (5) Diversity, non-discrimination and fairness; (6) Societal and environmental well-being; (7) Accountability. ALTAI is a method to drive the self-assessment and requires interdisciplinary expertise as well as a continuous evaluation procedure. In this section, we present and discuss our responses to the ALTAI questionnaire for each requirement in order to reflect the discussion and the concluding remarks of the workshop among the technical partners. We pursued an intermediate and a final workshop in which we used the ALTAI prototype web-based tool. The participants were representatives of the technical partners of the COALA consortium who had been developing the technological solution.

Human Agency and Oversight. AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. In this section, ALTAI asks to assess the AI system in terms of respect for human agency, as well as human oversight.

Human Autonomy. Human autonomy deals with the effect AI systems that are aimed at guiding, influencing or supporting humans in decision making processes. It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans and with the effect of AI systems on human affection, trust and (in)dependence. The COALA solution has been designed to interact, guide and support decisions by human end-users that affect humans. Through the DIA, it interacts with the operators, provides insights, supports the decisions as well as the training process of novice operators. The users know from the beginning that they interact with an AI system; thus, COALA could not generate confusion for the end-users on whether a decision, content, advice or outcome is the result of an AI algorithm. Moreover, COALA can potentially affect human autonomy by interfering with the end-user's decision-making process in an unintended way, since the outcomes of its functionalities are communicated to the user through the voice-enabled interface in a non-intrusive way. Although the risk is low, we have set up monitoring mechanisms in order to ensure that it will not cause over-reliance, since training is foreseen beforehand (e.g. the didactic concept and on-the-job training). In this sense, it is unlikely to manipulate human

behavior. We have also taken measures to mitigate the risk of manipulation by also protecting the technology stack of the solution.

Human Oversight. This subsection helps to self-assess necessary oversight measures through governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approaches. COALA is overseen by a mix of HITL and HIC approach. More specifically, the producer (manager or supervisor or administrator) controls how the employee should or can use the assistant. The user has control under these conditions. The manager has a complete control over the system. The employee (end user) has not a complete control but some degrees of freedom, depending also on their experience level. The humans that are involved in the use of the DIA have been given specific training on how to exercise oversight through the concept of AI-focused didactic concept [33]. We have also established detection and response mechanisms for undesirable adverse effects.

Technical Robustness and Safety. A crucial requirement for achieving TAI systems is their dependability and resilience. Technical robustness requires that AI systems are developed with a preventive approach against risks and that they behave reliably and as intended while minimising unintentional and unexpected harm. The questions in this section of ALTAI address four main issues: 1) security; 2) safety; 3) accuracy; and 4) reliability, fall-back plans and reproducibility. The COALA solution has not been certified for cybersecurity but it is compliant with various security standards that are inherent to the technology stack. This is achieved by the use of state-of-the-art technologies which are based on well-established technology standards. We use Keycloak in order to add authentication and secure services of the COALA components. Keycloak is based on standard protocols and provides support for OpenID Connect, OAuth 2.0, and SAML. In addition, COALA includes an anonymization component to fulfil privacy requirements. Therefore, it is not exposed to cyber-attacks, as it has also been verified by the IT departments of the use case partners. However, we have put measures in place to ensure its integrity, robustness and overall security. Finally, the users are informed about the duration of security coverage and updates.

General Safety. The COALA solution may potentially have adversarial, critical or damaging effects in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use. However, the probability of such cases is low because COALA does not automatically implement actions. The decision is finally taken by the operator. In each COALA business case, risks and risk levels have been defined from the very beginning of the project. These risks are continuously assessed throughout the evolution of the project and the progress of the development and deployment activities, and a specific process has been put in place in order to facilitate the consistent continuous risk assessment. Moreover, we have identified the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible resulting consequences. We have also assessed risks related to possible malicious use, misuse or inappropriate use as well as on the safety criticality levels of

their possible consequences. We have also assessed the dependency of critical system's decisions on its stable and reliable behavior.

Accuracy. COALA, as most of the AI systems, can potentially have critical, adversarial or damaging consequences. One cause of this can be a low level of accuracy which can lead to wrong guidance, misleading predictions, as well as inappropriate recommendations and advice. We have put in place measures to ensure that the data (including training data) used by the Data Analytics component is up to date, of high quality, complete and representative of the environment. This is also ensured by the data sources of the COALA use cases, as well as by the Data Management component which acquires and structures the data. We have also put in place a procedure to monitor and document COALA's accuracy by implementing mechanisms for evaluating the embedded algorithms and for providing interpretability insights. Further, these mechanisms consider whether the system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects. The results of the aforementioned accuracy evaluation mechanisms can be communicated to the end-users upon request either through the voice interface or through the Graphical User Interface (GUI).

Reliability, fall-back plans and reproducibility. COALA could potentially cause critical, adversarial or damaging consequences in case of low reliability and/or reproducibility. However, the probability of such cases is low. In any case, we have put in place procedures to monitor if the system meets the goals of the intended applications and whether specific contexts or conditions need to be taken into account to ensure reproducibility. We have also put in place verification and validation methods and documentation to evaluate and ensure different aspects of the system's reliability and reproducibility. Processes for the testing and verification of the reliability and reproducibility have been documented and operationalized. There have also been defined tested failsafe fallback plans to address COALA errors; they have been covered during the integration and have been validated in the context of the evaluation procedure. In addition, the Data Analytics component has embedded internal mechanisms for handling the cases where the system yields results with a low confidence score, while the voice interface has been subject to a UX study for chatbot breakdown assessment. All these activities were performed in close collaboration with the COALA use cases. COALA incorporates online continual learning in the sense of accommodating new knowledge while retaining previously learned experiences. In general, this is crucial for agents operating in changing environments and required to acquire, fine-tune, adapt, and transfer increasingly complex representations of knowledge. COALA tackles with this challenge in the following ways: (i) The didactic concept and the change management process teach and guide workers competencies when they collaborate with AI. It demonstrates how AI-specific worker education can help building trust in AI systems [33]; (ii) the Knowledge Management component captures best practices on the factory shop floor and facilitates knowledge acquisition, representation and inference [34]; (iii) the Data Analytics component incorporates ML algorithms that are capable of being updated, taking into account new data that are recorded [35].

Privacy and Data Governance. Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy. In COALA, we have considered the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection. Depending on the use case, we have established mechanisms that allow related flagging issues. There is the data anonymization service to protect the privacy of the workers and achieve General Data Protection Regulation (GDPR) compliance. In addition, in order to thoroughly implement the GDPR, we have defined a Data Protection Officer (DPO) role in the consortium from the very beginning of the project so that he is involved in all the phases of the COALA lifecycle. We have also adopted measures to enhance privacy by design and default (e.g. encryption, anonymisation), which are continuously assessed throughout the development phases. It should be noted that COALA does not use or process personal data (including special categories of personal data) when being trained and developed. Where applicable, we have adopted a policy for data minimization. We have also taken into account the right to withdraw consent, the right to object and the right to be forgotten in the COALA solution. We have considered the privacy and data protection implications of data collected, generated or processed as well as the privacy and data protection implications of non-personal training-data or other processed non-personal data.

Transparency. A crucial factor for achieving TAI is transparency which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system. Technical robustness requires that AI systems be developed with a preventive approach to risks and in a way that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured.

Traceability. This subsection helps to self-assess whether the processes of the development of the AI system, i.e. the data and processes that yield the AI system's decisions, is properly documented to allow traceability, increase transparency and, build trust in AI in society. In COALA, we have put in place measures to continuously assess the quality of the input data to the AI system.

Explainability. This subsection helps to self-assess the explainability of the AI system. The questions refer to the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes. Explainability is crucial for building and maintaining users' trust. AI-driven decisions should be explained and understood to those directly and indirectly affected, in order to allow contesting of such decisions. An explanation as to why a model has generated

a particular output or decision is not always possible. These cases are referred as “black boxes” and require particular attention. In those circumstances, other explainability measures may be required, provided that the AI system as a whole respects fundamental rights. The degree to which explainability is needed depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human lives. COALA explains the decisions and all its outcomes to the users through an explainability engine. Moreover, it incorporates a Large Language Model (LLM) which segments from a pdf a recommendation allowing the user going back where the answer comes from. COALA continuously surveys the users to ask them whether they understand the decisions of the AI system taking, at the same time, into account that the operator should not be disturbed by unnecessary detailed explanations.

Communication. This subsection helps to self-assess whether the AI system’s capabilities and limitations have been communicated to the users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy as well as its limitations. Since COALA is based on chatbot and voice-enabled technologies, in order to facilitate the interaction between humans and AI, the users are explicitly informed that they interact with an AI system and not with a human. We have established mechanisms to inform users about the purpose, criteria and limitations of the decisions generated by COALA. To do this, we use a "capabilities" intent explaining what the assistant can do. We use a "FAQ" intent pointing users at a learning nugget with basics about digital assistants. We use an "out of scope" intent to indicate when the assistant cannot answer/help the user (because training data contains example utterances that are out of scope). Regarding the LLM in particular, if a question is out of context (i.e. if semantic similarity is below a threshold), it does not provide answers. We communicated the technical limitations and potential risks of the AI system to end-users, since the results about the level of accuracy and/or error rates are available to be exposed to the user upon request. Moreover, we provided appropriate training material and disclaimers to users on how to adequately use the COALA system as part of a didactic concept. The evaluation of the COALA solution incorporated user tests for the aforementioned findings.

Diversity, Non-discrimination and Fairness. In order to achieve TAI, inclusion and diversity should be enabled throughout the entire AI system’s life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models. Identifiable and discriminatory bias should be removed in the collection phase where possible. AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance.

Avoidance of unfair bias. In COALA, we have established a set of procedures to avoid creating or reinforcing unfair bias, both regarding the use of input data as well as for the algorithm design [37]. We have considered diversity and representativeness of end-

users and subjects in the data. We research and use mostly open-source state-of-the-art technical tools in order to improve understanding of the data, model and performance. We assessed and put in place processes to test and monitor for potential biases during the entire lifecycle of COALA (e.g. biases due to possible limitations stemming from the composition of the used data sets - lack of diversity, non-representativeness). We have put in place educational and awareness initiatives to help system designers and developers be more aware of the possible bias they can inject in designing and developing the AI system. In this context, we have created an Ethics Board and we performed an ethics survey. Depending on the use case, we ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance. Moreover, we have established ways of communicating on how and to whom such issues can be raised yet, while we have also identified the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users. In COALA, we use the widely used generic definition of "fairness" without having elaborated on its instantiation to the requirements of the project. However, the COALA ethics survey provides the means to do this.

Accessibility and universal design. Particularly in business-to-consumer domains, AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for people with disabilities, which are present in all societal groups, is of particular importance. In this context, we have ensured that Universal Design principles are taken into account during every step of the planning and development process, while we have also taken into account the impact on the potential end-users. We have also assessed whether the team of developers are engaged with the possible target end-users. The COALA project, as a European collaborative project, dictates the close collaboration between the technical partners and the use cases by its nature. Therefore, following an agile software development methodology, the use case partners were continuously interacting with the technical partners. The rest of the ALTAI questions are not applicable to the COALA solution; however, there is not such an option in the alternative responses. Consequently, our responses were given in a way to maintain their neutrality and to avoid affecting the resulting ALTAI recommendations to the degree this is feasible.

Stakeholder participation. In order to develop TAI, it is advisable to consult stakeholders who may directly or indirectly be affected by the AI system throughout its life cycle. COALA, as a European collaborative research project, by nature includes the close collaboration between the use case partners and the technical partners in its design and development. In addition, the project included tasks dedicated to ethics and human-centric AI, while its governance included an Ethics Board. However, the stakeholders were mainly high-skilled industrials and developers, something which is not a representative case in AI software development.

Societal and Environmental Wellbeing. In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment

should be considered as stakeholders throughout the AI system's life cycle. Ubiquitous exposure to social AI systems in all areas of our lives may alter our conception of social agency, or negatively impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could equally affect peoples' physical and mental well-being. The effects of AI systems must therefore be carefully monitored and considered. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals.

Environmental Wellbeing. This subsection helps to self-assess the positive and negative impacts of the AI system on the environment. Measures to secure the environmental friendliness of an AI system's entire supply chain should be encouraged. COALA has not any potential negative impacts system on the environment taking into account the Context of the ALTAI questions. In this sense, the respective questions of the ALTAI questionnaire in this section were not applicable. We answered positively in order not to affect the resulting recommendations.

Impact on work and skills. This subsection helps self-assess the impact of the AI system and its use in a working environment on workers, the relationship between workers and employers, and on skills. The COALA solution has a major impact on human work and work arrangements, since it is aimed at supporting the operations on the shop floor in manufacturing environments. Furthermore, COALA adopts the didactic concept as well as learning nuggets in order to support a training approach for introducing changes through the advice of dialogs by the voice interface [33]. This approach results in measurable changes in workers behavior. Moreover, COALA implements an assistant function to support novice workers in their learning and working activities while reconfiguring and operating production lines [15,34]. In order to prepare the implementation, deployment, and evaluation of the COALA concepts and technologies, we have informed and consulted the impacted workers and their representatives in advance in order to ensure that the work impacts are well understood. This is also part of the change management procedure that we have adopted [37]. We use a didactic concept to teach workers about challenges, capabilities, and risks of DIAs. We use a change management process focused on AI to prepare managers and workers for human-AI collaboration. To that extent, we have taken measures to counteract de-skilling risks. Workers keep focusing on knowledge-intensive tasks while the assistant takes over repetitive tasks (that are subject of de-skilling).

Accountability. The principle of accountability requires that mechanisms are put in place to ensure responsibility for the development, deployment and/or use of AI systems. This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties.

Auditability. This subsection helps to self-assess the existing or necessary level that would be required for an evaluation of the AI system by internal and external auditors. In applications affecting fundamental rights, including safety-critical applications, AI

systems should be able to be independently audited. We have established mechanisms that facilitate the AI system's auditability.

Risk Management. Both the ability to report on actions or decisions that contribute to the AI system's outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. We have established an "AI ethics review board" to discuss the overall accountability and ethics practices, including potential unclear grey areas. We have also established processes for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) or workers to report potential vulnerabilities, risks or biases in the AI system. In addition, we have not put in place redress by design mechanisms in cases COALA significantly adversely affect individuals.

4.2 Results and Recommendations

The outcomes of the ALTAI tool are: (i) A visualisation of the self-assessed level of adherence of the AI system with the 7 requirements for TAI; and, (ii) Recommendations based on the answers to the questionnaire. **Fig. 1** depicts the results of the Assessment List in the form of a Polar diagram for the 7 requirements of ALTAI. **Table 3** presents the resulting recommendations per ALTAI requirement. It should be noted that some questions are not applicable to the COALA solution or the alternative responses that are provided do not represent accurately the opinion of the COALA consortium. These issues inevitably affect the scores for the 7 ALTAI requirements. This fact dictates the addition of one more step in this self-assessment procedure, a validation of the resulting recommendations with regards to the scope of the AI system under examination.

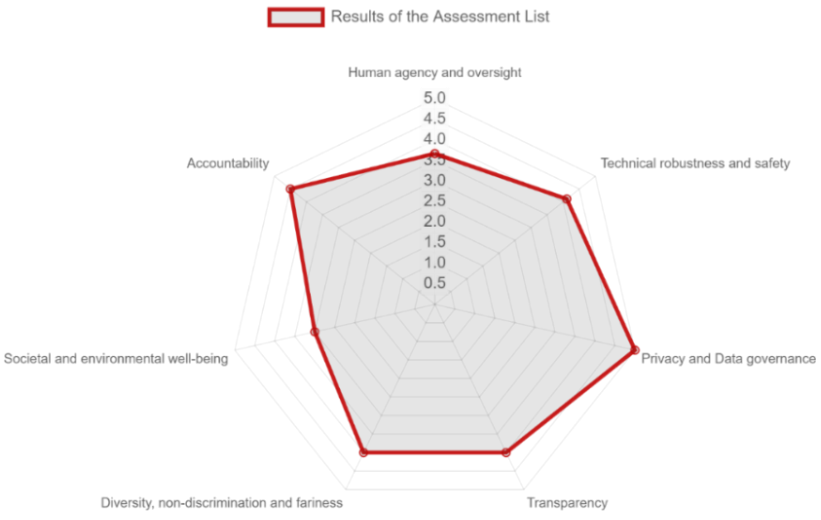

Fig. 1. The resulting score of ALTAI for the 7 requirements.

Table 3. ALTAI recommendations per requirement.

Resulting ALTAI Recommendations for each out of the 7 Requirements	
1	**Human agency and oversight**
No recommendation for this requirement.	
2	**Technical robustness and safety**
No recommendation for this requirement.	
3	**Privacy and Data Governance**
i.	Whenever possible and relevant, align the AI-system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for data management and governance.
4	**Transparency**
No recommendation for this requirement.	
5	**Diversity, non-discrimination and fairness**
i.	Your definition of fairness should be commonly used and should be implemented in any phase of the process of setting up the AI system.
ii.	Consider other definitions of fairness before choosing one.
iii.	Consult with the impacted communities about the correct definition of fairness.
iv.	Ensure a quantitative analysis to measure and test the applied definition of fairness.
v.	Establish mechanisms to ensure fairness in your AI system.
vi.	You should assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion.
vii.	You should assess the risk of the possible unfairness onto the end-user's communities.
6	**Societal and environmental well-being**
No recommendation for this requirement.	
7	**Accountability**
i.	3rd party auditing can contribute to generate trust in the technology and the product itself. Additionally, it is a strong indication of adhering to industrial standards.
ii.	If AI systems are increasingly used for decision support or for taking decisions themselves, it has to be made sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this.
iii.	A risk management process should include new findings since initial assumptions about the likelihood of occurrence for a specific risk might be faulty.
iv.	Acknowledging that redress is needed when incorrect predictions can cause adverse impacts to individuals is key to ensure trust.

5 Conclusions and Remarks on the ALTAI Tool

In this paper, we presented ALTAI's application on a DIA for the Operator 5.0, but we also conclude to some remarks:

1. ALTAI has been designed for end products; it does not address the various phases of software development lifecycle. This could ensure the compliance of the AI system with ethics guidelines during its development, but also enable early

improvements in the context of agile software development. Therefore, there is the need to create ALTAI variants for the various development phases or to connect some questions to different phases to treat them differently (e.g. reducing their scoring weight). In addition, the assessment could outline potential risks if the system developers do not address a shortcoming.
2. ALTAI incorporates generic questions aiming at addressing every AI system. However, for example, in the case of COALA, the AI system refers to a manufacturing environment, i.e. a professional environment with expert and qualified users. In contrast, an AI system referring to a different application domain or even more a generic audience of end-users, may have different requirements. Therefore, there is the need for domain-specific adaptations of ALTAI with context-specific questions.
3. ALTAI considers as "AI system" the software and does not treat it as a socio-technical system, potentially leading to disregard of unforeseen challenges. Moreover, the way artificial agents learn may not be understandable to humans, making uncertainty and unpredictability present to a higher degree than in traditional systems.
4. No alternative response is accurate for some questions. Some responses should have been "not applicable", "not yet", or "to some extent" instead of the options "yes / no / don't know". Given the limitation (1), options such as "not yet" would indicate that something has not been implemented yet, but it has been planned. Given the limitation (2), ALTAI in each current form could have provided options such as "not applicable". These affect the resulting assessment score and the recommendations, some of which may not be applicable. To this end, these results need further validation during the development activities. In order to provide different response options, ALTAI could adopt a Likert-scale approach.
5. While ALTAI covers the seven key requirements of TAI, the current structure hinders its applicability. Several sections of ALTAI are unbalanced, since some of them cover large parts of the assessment list, while there are overlapping and redundant questions. ALTAI could be reorganized to support readability.
6. In the long-term, the tool could be enhanced with generative AI, e.g. generating stories about a hypothetical market introduction of the AI system with a shortcoming or good rating resulting in story part to explain it. E.g., a shortcoming in "privacy and data governance" could lead to a story part where personal data is misused.

Acknowledgements. This work is funded by the European Union's H2020 project COALA (https://www.coala-h2020.eu/) (Grant agreement No 957296). The work presented here reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

References

1. AI HLEG.: Ethics Guidelines for Trustworthy AI. Brussels: European Commission (2019)
2. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Vayena, E.: AI4People-An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines **28**(4), 689-707 (2018)

3. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Goldenberg, A.: Do no harm: a roadmap for responsible machine learning for health care. Nature medicine **25**(9), 1337-1340 (2019)
4. Baneres, D., Guerrero-Roldán, A. E., Rodríguez-González, M. E., Karadeniz, A.: A Predictive Analytics Infrastructure to Support a Trustworthy Early Warning System. Applied Sciences **11**(13), 5781 (2021)
5. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion **76**(1), 89-106 (2021)
6. Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence. Electronic Markets **31**(2), 447-464 (2021)
7. Georgieva, I., Lazo, C., Timan, T., van Veenstra, A. F.: From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI and Ethics, 1-15 (2022)
8. Kazim, E., Koshiyama, A.: AI assurance processes. Available at SSRN 3685087 (2020).
9. Smuha, N. A.: The EU approach to ethics guidelines for trustworthy artificial intelligence. Computer Law Review International **20**(4), 97-106 (2019)
10. Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., Bouarfa, L., Chatila, R., Coeckelbergh, M., Dignum, V., Van Wynsberghe, A.: The assessment list for trustworthy artificial intelligence (ALTAI). European Commission (2020)
11. Borg, M., Bronson, J., Christensson, L., Olsson, F., Lennartsson, O., Sonnsjö, E., Ebabi, H., Karsberg, M.: Exploring the Assessment List for Trustworthy AI in the Context of Advanced Driver-Assistance Systems. In: 2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics), pp. 5-12. IEEE (2021)
12. Romero, D., Stahre, J.: Towards the resilient operator 5.0: The future of work in smart resilient manufacturing systems. Procedia CIRP, 104, 1089-1094 (2021).
13. Freire, S.K., Niforatos, E., Wang, C., Ruiz-Arenas, S., Foosherian, M., Wellsandt, S., Bozzon, A.: Lessons learned from designing and evaluating CLAICA: a continuously learning ai cognitive assistant. In: Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 553-568 (2023)
14. Bousdekis, A., Wellsandt, S., Bosani, E., Lepenioti, K., Apostolou, D., Hribernik, K., & Mentzas, G. Human-AI collaboration in quality control with augmented manufacturing analytics. In: Advances in Production Management Systems IFIP WG 5.7 International Conference, France, pp. 303-310. Springer International Publishing (2021)
15. COALA project - Deliverable 2.5. Digital intelligent assistant core for manufacturing demonstrator– version 2. Available online: https://ncld.ips.biba.uni-bremen.de/s/S8W8dmm2ei4RbGw (last accessed: 03/06/2024)
16. Future of Life Institute: Asilomar AI Principles (2017). Available online: https://futureoflife.org/ai-principles/ (last time accessed: 14/3/2024)
17. Université de Montréal: Montreal Declaration for a Responsible Development of AI (2017). Available online: https://www.montrealdeclaration-responsibleai.com/the-declaration (last time accessed: 14/3/2024)
18. Chinese National Governance Committee for the New Generation Artificial Intelligence.: Governance Principles for the New Generation Artificial Intelligence–Developing Responsible Artificial Intelligence (2019). Available online: https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html (last time accessed: 14/3/2024)
19. UK House of Lords. AI in the UK: ready, willing and able? (2017). Available online: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm (last time accessed: 14/3/2024)

20. OECD: OECD Principles on AI (2019). Available online: https://www.oecd.org/going-digital/ai/principles/ (last time accessed: 14/3/2024)
21. Metzinger, T.: Ethics washing made in Europe. Der Tagesspiegel (2019)
22. Veale, M.: A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. European Journal of Risk Regulation **11**(1), (2020)
23. European Commission: Proposal for a Regulation on a European approach for Artificial Intelligence (No. COM (2021) 206 final). European Commission, Brussels (2021).
24. Stix, C.: The ghost of AI governance past, present and future: AI governance in the European Union. arXiv preprint arXiv:2107.14099 (2021)
25. AI Now Institute: Algorithmic impact assessments: a practical framework for public agency accountability (2018). Reisman, D., Schultz, J., Crawford, K., Whittaker, M.: Algorithmic impact assessments: A practical framework for public agency. AI Now **9** (2018)
26. IEEE: IEEE 7010-2020—IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being (Standard). IEEE (2020)
27. Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., Westerlund, M.: Z-Inspection®: a process to assess trustworthy AI. IEEE Transactions on Technology and Society **2**(2), 83-97 (2021).
28. Stahl, B. C., Leach, T.: Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: An empirical test of the European Union Assessment List for Trustworthy AI (ALTAI). AI and Ethics **3**(3), 745-767 (2023)
29. Stefani, T., Deligiannaki, F., Berro, C., Jameel, M., Hunger, R., Bruder, C., Krüger, T.: Applying the Assessment List for Trustworthy Artificial Intelligence on the development of AI supported Air Traffic Controller Operations. In: 2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC), pp. 1-9. IEEE (2023)
30. Rajamäki, J., Gioulekas, F., Rocha, P. A. L., Garcia, X. D. T., Ofem, P., Tyni, J.: ALTAI Tool for Assessing AI-Based Technologies: Lessons Learned and Recommendations from SHAPES Pilots. Healthcare **11**(10), 1454, MDPI (2023)
31. Radclyffe, C., Ribeiro, M., Wortham, R. H.: The assessment list for trustworthy artificial intelligence: A review and recommendations. Frontiers in Artificial Intelligence **6**(1), 1020592 (2023)
32. Gavornik, A., Podroužek, J., Mesarcik, M., Solarova, S., Oresko, S., Bielikova, M.: Utilising the Assessment List for Trustworthy AI: Three Areas of Improvement. ceur-ws.org (2022)
33. COALA project - Deliverable 3.6. AI-focused Didactic Concept for Factory Workers – final. Available online: https://ncld.ips.biba.uni-bremen.de/s/Wy7ywFBFjLw8oKx (last accessed: 03/06/2024)
34. COALA project - Deliverable 3.4. Cognitive Advisor Service – Version 2. Available online: https://ncld.ips.biba.uni-bremen.de/s/LkgM8BLiMmmgn7q (last accessed: 03/06/2024)
35. Fikardos, M., Lepenioti, K., Bousdekis, A., Bosani, E., Apostolou, D., Mentzas, G.: An Automated Machine Learning Framework for Predictive Analytics in Quality Control. In: IFIP International Conference on Advances in Production Management Systems (pp. 19-26). Cham: Springer Nature Switzerland (2022).
36. COALA project - Deliverable 7.6. Report on the application of ethical principles for AI in manufacturing – Final. Available online: https://ncld.ips.biba.uni-bremen.de/s/9PQPpMTgSpHrtqQ (last accessed: 03/06/2024)
37. COALA project - Deliverable 5.3. Change management process for human – AI Collaboration – Final. Available online: https://ncld.ips.biba.uni-bremen.de/s/s6nRtrNbGKsTppa (last accessed: 03/06/2024)