POLICY BRIEF

ETAPAS

ETAPAS and ALTAI: two European Trustworthy AI assessment methodologies

How to ensure an effective trustworthiness assessment journey to your AI application using the most appropriate methodology

Relevant for:

- Policy Makers
- Innovators within the Public Sector
- Professionals within the Public Sector
- Private companies that offer services to Public Organisations

Table of Contents

Figures

Tables

Executive Summary:

This Policy Brief entitled “ETAPAS and ALTAI: two European Trustworthy AI assessment methodologies” aims at analysing the main characteristics of two different trustworthy AI assessment tools, namely the ETAPAS Responsible Disruptive Technology (RDT) Framework and methodology for the ethical social and legal assessment of Disruptive Technologies in the Public Sector and ALTAI “Assessment List for Trustworthy AI”.

After an initial description of the main features presented by the two assessment tools, a comparative analysis is carried out in order to determine and recognise when it is more appropriate to use one tool than the other, reflecting on the main advantages the ETAPAS RDT Framework and ALTAI present as well as highlighting the existing differences, to then identify specific use-cases for both tools as highlighted in the paragraph “Lessons learned from the comparative analysis”.

Finally, a set of recommendations is formulated to encourage trustworthy AI adoption and improve the ethical assessment of AI in the public sector. Such recommendations briefly provide for:

- Appling the assessment process to all Disruptive Technologies and tailoring it to the different sectors;
- Starting the ethical, social, and legal assessment from the design phase of the application and throughout the application’s entire lifecycle;
- Involving all relevant stakeholders in the assessment process, including users and final users;
- Ensuring reliable results;
- Monitoring results over time.

1. Introduction

This document is aimed at understanding how AI can be safely applied in the Public Sector so to set up Trustworthy systems. By doing so, Public Administrations (PAs) can efficiently gather huge amounts of information, automate repetitive tasks, modernise their communication management systems, ensure higher levels of security, and improve their overall efficiency. This being true, embedding AI in such context is an operation which is not free from risks, as unethical behaviour might be adopted thus damaging the local community served by PAs. Therefore, one should adopt preventive measures to contrast and tackle the incorrect use of AI applications by involving both users and the network of stakeholders which will be then impacted by the consequences such behaviour would generate. These measures can be concretely taken by running an assessment aimed at evaluating one’s compliance with ethical requirements identified as key pillars to build a Trustworthy environment in which AI is perfectly integrated. Such opportunity is indeed offered by the ALTAI self-assessment tool by the High-Level Expert Group on AI of the European Commission, as well as by the innovative governance platform developed by the H2020-funded project ETAPAS, targeting specifically PAs.

The purpose of this Policy Brief is precisely to present these two very useful tools, highlighting their characteristics and vocations and comparing them to derive useful policy recommendations to improve the ethical assessment of AI in the public sector.

To this end, the document will first introduce both the ETAPAS Project and the Assessment List on Trustworthy AI, namely ALTAI, exploring their assessment methodologies as well as the feedbacks gathered on them from their users, with a comparative analysis aimed at underlining

the main characteristics of both tools, including strengths, weaknesses, and areas for improvement. At the end of this analysis, we will produce policy recommendations to facilitate trustworthy AI assessment for the benefits of PAs and citizens.

2. The ETAPAS Project

ETAPAS – Ethical Technology Adoption in Public Administration Services – is a project funded by the European Commission through the Horizon2020 programme and led by the Italian Ministry of Economy and Finance (MEF). The idea arose from the need, especially felt in the public sector, to ensure the ethical, social, and legal compliance of the adoption of disruptive technologies.

A disruptive technology is an innovation that significantly alters the way that consumers, industries, or businesses operate, while displacing a well-established product or technology and creating a new industry or market. The project focuses on three of these technologies, deemed particularly complex in terms of ethical impact: AI, Robotics and Big & Open Data.

Above all, the project aims to co-design and validate with the public administrations involved in the broad consortium, composed by 14 organisations from 8 European countries, an assessment framework and tailoring methodology specific to the public sector.

Assessment framework and methodology are tested and refined in the course of the project on four real use cases:

- *Ethically Responsible Big Open Data*, the MEF use case on the human resources management platform most used by Italian public administrations “NoiPA”
- *Robot-mediated rehabilitation*, the IIT-FDG use case focusing on robots used for the assessment of patients’ walking abilities
- *Municipality chatbot – Kari*, the SINTEF-Prokom use case concerning the Kari chatbot used by many municipalities in Norway to inform citizens about municipal services and its possible integration with the national NAV chatbot
- *Public Organizations Multi-Factor Misinformation Handling*, the use case form CERTH and the Municipality of Katerini, addressing the deployment of AI for fake news detection and prioritization of emerging issues in the municipality.

The selected use cases employ different technologies and have different maturity to achieve a more comprehensive validation of the framework.

Upon completion of validation activities, the final output will be a prototype software platform and a governance model that will enable public sector organisations to easily assess and manage ethical, social, and legal risks and impacts throughout the entire technology life cycle.

2.1 The Responsible Disruptive Technology Framework

The ETAPAS assessment framework is called RDT – Responsible Disruptive Technology – framework and consists of the four components shown in Figure 1 below.

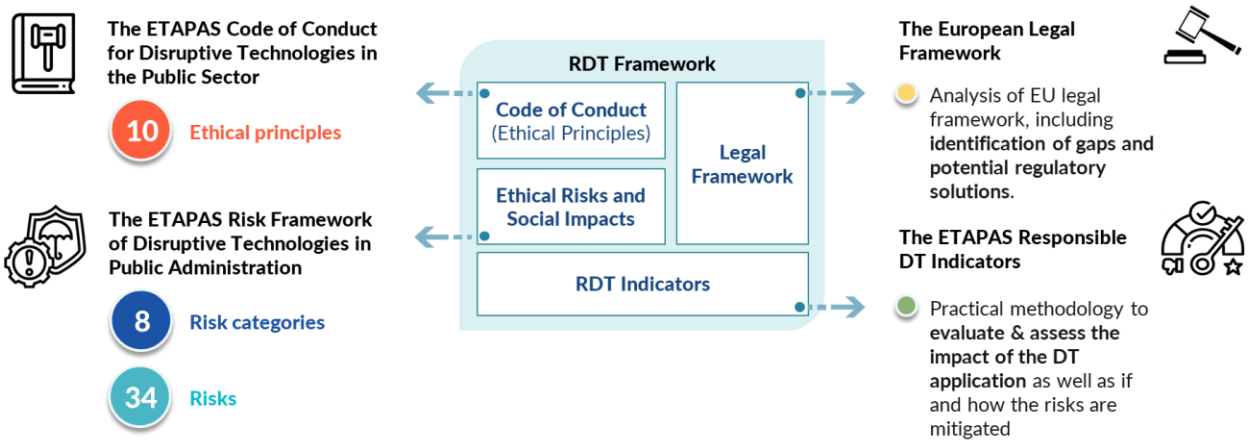

Figure 1 - RDT framework

The **generic Code of Conduct for the adoption of Disruptive Technologies (DTs) in the Public Sector** sets out 10 ethical principles to ensure compliance with by Public Sector organisations that are adopting DTs, namely **Environmental Sustainability**, **Justice, equality, and the rule of law**, **Transparency and explainability**, **Responsibility and accountability**, **Safety and security**, **Privacy**, **Building an ethical culture involving employees**, **Retaining human contacts**, **Ethical public-private cooperation**, and **Continuous evaluation and improvement**.

Moreover, the Code of Conduct is generic as it is not designed to be employed for a specific technology or for a specific application area, therefore a co-design process with all the relevant stakeholders within each organisation is suggested to further detail and tailor it.

The Legal Framework, provides an analysis of all the referring **European legal framework**, including European guidelines and the AI Act.

The overview of the ethical risks and social impacts of DTs, includes an **extensive mapping of risks and impacts** that Public Sector organisations face when adopting or using DTs. To this extent, examples of significant risks that can be identified are represented by **Replacement of human agency** and **Exclusion of individuals** among the ***Risks concerning direct interaction with humans***, **Illegal behaviour** – under the umbrella of ***Legal Risks*** -, and finally **Lack of transparency and explainability**, **Relations with the private sector**, and **Unclear accountability and responsibilities** within the ***Governance Risks*****.**

The **RDT indicators** to practically measure those risks and impacts, include both qualitative and quantitative indicators that will be **monitored through the ETAPAS governance platform** prototype.

For the co-design of the framework, the consortium used a **participatory and iterative methodology**. Bilateral meetings, co-design workshops, plenary meetings, internal consultations via surveys and public consultations on the ETAPAS website were organised. In addition, ten

main best practices[1] of ethical assessment were analysed and served as the basis for the model, including ALTAI.

As for the RDT indicators framework, it includes two types of indicators:

1. **assessment indicators** structured as a fact-based general checklist and that may be text, numeric, single choice, multiple choice, yes/no, ratio grid and direct feedback indicators; and
2. **computational indicators** tailored to each specific use case/DTA. The analysis of the best practices revealed the limitations of purely quantitative approaches to this type of assessment concerning especially social-ethical issues and showed the merits of a check-list approach. Most of the analysed framework4 were structured as a list of questions categorised according to specific criteria, such as the relevant ethical issue or the risk to be avoided or mitigated.

Our approach has been to develop the assessment indicators from individual risks to ensure their coverage and to track the indicators towards both the main risk category and the relevant ethical principle to allow categorisation. We later determined that the most efficient categorisation criterion was risk categories, as they were more specific to the indicators than the ethical principles, whereas individual risk could not be applied as a criterion because indicators almost always measure more than one risk.

Moreover, many of the frameworks analysed included many self-assessment questions, which are not very objective and generalisable. Given the stringent socio-ethical requirements of the public sector and the ambition to generalise the framework, we chose a **fact-based approach**. This type of approach consists of asking the person conducting the assessment whether or not they have carried out an action, whether the technological solution used has certain characteristics, or something that is factual and does not require subjective judgement. Questions that are asked using this approach are usually structured in this way: “Has your organisation adopted a set of ethical principles/code of conduct concerning the adoption of technologies and shared it with all employees?”, “Did your organisation establish a continuous chain of responsibility for all roles involved in the design and implementation lifecycle of the use case?”, “Did your organisation put in place the relevant procedures to ensure that the data collection devices do not expose the DTA to cybersecurity threats?”, “Did you perform active monitoring, review and regular tuning when appropriate on the DTA?”, “Did the organization put in place ways to measure whether its system is making an unacceptable number of inaccurate predictions?”, “Did your organization conduct an impact assessment (e.g. probability and/or severity of harm) on individuals and organizations who are affected by the DTA?”.

Some of the assessment indicators are classified as “**direct feedback” indicators**. By “direct feedback” indicators we mean those types of assessment question that require direct answers from the final user, defined as “the person that is impacted by the decision taken by the DT, without directly interacting with it.” These are the only indicators that involve subjective self-assessment. A direct feedback indicator is for example: “How much does the user feel safe while interacting with the DTA? [Very safe, somehow safe, not totally safe, unsafe]”. These types of

[1] The UK Gov AI ethics safety guidance & The ICO framework, WEF AI Governance framework, ALTAI, UK Gov Data Ethics framework, NOREA, UK Gov A guide to using artificial intelligence in the public sector, The ICO AI and Data Protection Risk Toolkit, ROBIA, ROBIA, NIST Framework for Improving Critical Infrastructure Cybersecurity

indicators are particularly useful to assess qualitative aspects of the final user experience with the DTA.

By “**computational” indicators**, on the other hand, we mean those metrics that can be used to derive automatically from a set of data an assessment score for a certain risk-aspect of the DTA. These indicators, due to their special automatic nature, need to be tailored to each specific use case and deserve a separate sheet of the framework. An example of a “computational” indicator for the use case concerning Robot-mediated rehabilitation is: “Event from navigation system recorded online on the platform to measure potential physical harm and collision with the patient/therapist.” This type of indicator allows certain aspects of the DTA to be continuously monitored, making it possible to monitor particularly risky areas in relation to the specific use case.

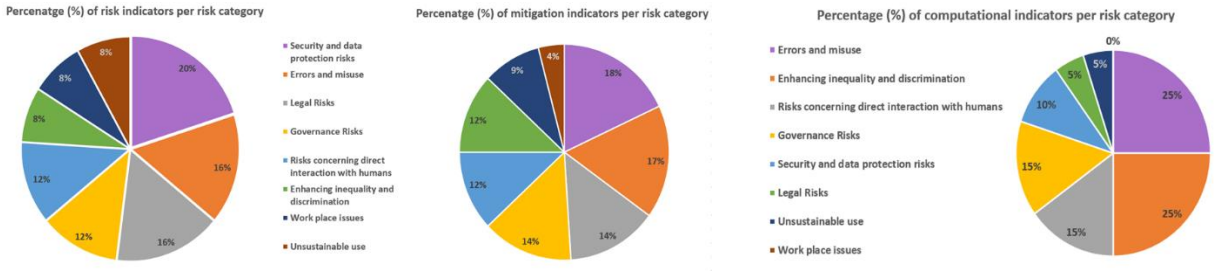

Figure 2 - ETAPAS RDT indicators by type

Furthermore, to ensure **adherence to needs and completeness of the framework**, for each indicator the following key dimensions have been defined:

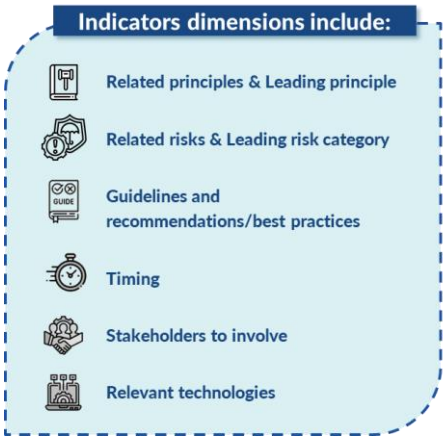

Figure 3 - Indicators dimensions

Indicators are configured in three different types of assessment within the prototype ETAPAS governance platform:

1. **pre-development assessment** – to be filled-in before implementation, during the design phase of the DT life cycle.
2. **during implementation assessment** – to be filled-in during the implementation of the DT solution

3. **post-development assessment** – to be filled-in during the deployment phase

The suggested approach is to perform periodically the assessment, by filling in the relevant assessment questionnaire on the ETAPAS governance platform approximately every six months, starting from the design phase. On the other hand, for computational indicators, once set up and connected with the platform, they are monitored constantly.

Results are given in the form of three dashboard:

- **Risk mitigation results**, showing both the **risk score** – assessing the relevance of the risk to the specific DT application – and the **mitigation score** – assessing the effectiveness of any implemented actions in mitigating the risk – resulted from the assessment questionnaire. These scores are shown both for the whole assessment that by individual risk category. Furthermore, by clicking on the risk categories' results, tailored recommendations are shown to the users suggesting actions to be put in place to improve his own results and his use case ethical, social and legal compliance.
- **Trend** dashboard, showing the behaviour of the use case through time, thus trend in scores both overall risk and mitigation scores and per each risk category.
- **Computational** dashboard, showing the results got from computational indicators. This dashboard has two sections, a first showing text results and a second showing numeric results and related averages and trends. It can also be filtered by risk category to explore specific risk aspects of the use case.

In addition, all results can be printed as pdf files.

2.2 Tailoring and Validation methodology

The validation methodology is thought to help Public Administrations during the DT adoption process in **analysing the context and consequently tailor the RDT Framework.** Just as the code of conduct is designed to be **customised to the sector and type of technology of the specific use case**, the entire framework and assessment can be tailored following the methodology that will be provided in the **ETAPAs Governance model for PAs**.

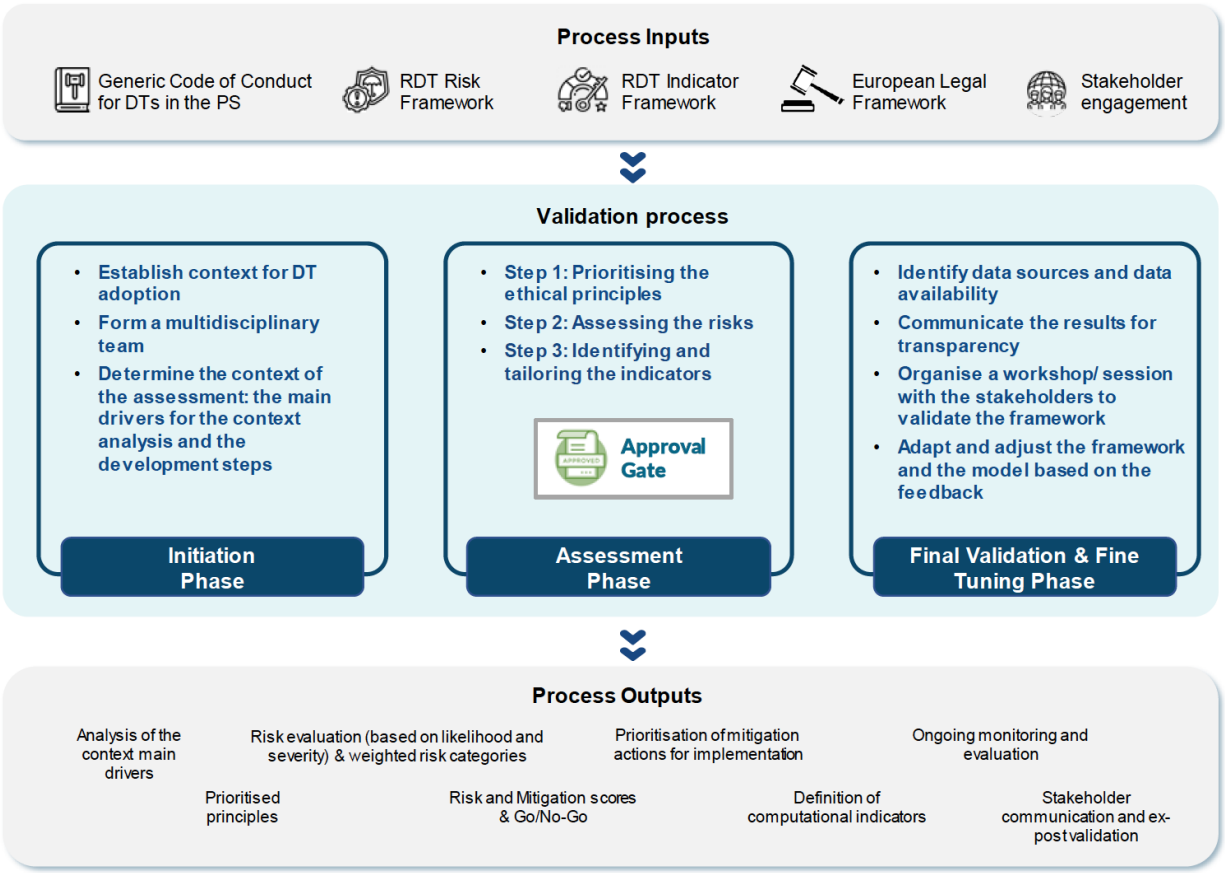

Figure 4 - Tailoring and validation methodology

As showed by Figure 4, the methodology uses as process inputs all the elements from the **ETAPAS RDT framework** as well as **stakeholder engagement techniques**.

The **Validation Process** itself, is structured in three phases:

- **The Initiation Phase –** where a multidisciplinary team is built-up following specific guidelines and then it analyses the context of the use case. As a result, the team determines the **context drivers** for the specific use cases, upon which ethical principles and risk prioritisation will be based**.** To jointly determine these results, an **interactive workshop** is organized.
- **The Assessment Phase –** constitutes the core of the assessment. It is made-up of three key steps, namely:
    - **Step 1: prioritise the ethical principles**
    - **Step 2: assessing the risks**
      In these two steps, based on the context analysis, the team defines a prioritisation for the ethical principles and for the risk and risk categories of the ETAPAS RDT framework. To jointly determine these results (Initiation phase, Step 1 and Step 2), one or two **interactive workshops** are organised.
    - **Step 3: identifying and tailoring the indicators**
      In this phase, the team should perform a **pre-development assessment questionnaire**, which is a fact-based checklist-type questionnaire, whose score

are tailored based on the prioritisations and analysis carried out in the previous steps. Also, the use case owner should ensure the **set-up of the DT solution**, so that **relevant events can be sent** directly to the **ETAPAS governance platform** and monitored as **computational indicators**. These indicators are defined ad-hoc based on the specific use case.

This phase leads to the **approval gate**, where a risk and a risk mitigation score are provided for the specific use case, and **the use case owner might be aware of whether its DT solution is acceptably safe or not**. Accompanying this response, he will also be informed about **risk-related mitigation actions and best practices he might be following in order to improve the DT adoption process**. Again, it is suggested to perform this assessment at the design stage and then perform it periodically on a semi-annual basis.

- **The Final Validation and Fine-tuning Phase** - consists of implementing mitigation actions to ensure social, legal and ethical compliance and lower risk levels, including sharing with relevant stakeholders and the community at large DT's progress for transparency and continuing to monitor and improve its performance.

The methodology was co-designed with public sector stakeholders and is being tested and validated with the four ETAPAS use cases.

2.3 Feedbacks and lessons learned from use cases

What has emerged so far through the implementation of the ETAPAS use cases and the concrete testing of the co-designed methodologies is that the developed methodology is very useful and complete but needs the involvement of relevant stakeholders and support in order to be implemented quickly and in the most effective way. The use case teams had different compositions and technologies at different maturity levels. Remarkably, lower maturity UC teams found it easier to think about the future and anticipate possible risks as well as solutions, while higher maturity UC teams found it more difficult to mentally get out of the current experiment configuration. Unsurprisingly, it proved particularly useful for the effectiveness of the workshops to have DT users in the team. **This early data confirms the need to undertake this initial assessment prior to development and with a team involving users of the technology.** All use cases had contextual aspects to be highlighted in order to better assess potential risks and mitigation actions to be implemented. This highlights the **need to analyse the context** of technology adoption. It is worth noting that almost all UC teams have **weighed heavily "Legal risks"**, being this class of risks very important for **public sector organisations**. In addition, many of them weighed heavily the risk category "Workplace Issues" probably fearing a lack of the required skills in their organisations. For all use cases the weights were distributed fairly evenly, without excessively heavy or light risk categories, reminding us that **all risk categories are to be monitored**.

Finally, ETAPAS use cases also showed how, despite the tool's completeness, there is still a general lack of clarity when completing the assessment, a weakness that is now getting tackled especially by working on the platform, which will also allow to estimate the costs related to such operations. Simultaneously, use cases highlighted the need of a structured process, based on our tested proposal, within the entities that complete the assessment, as the assessment itself requires quite long time to be carried out appropriately.

Presenting the framework and methodology to the students of the **AI4GOV Master**, a project co-funded by the European Commission, they mostly agreed on the need for a **structured methodology of interaction** between relevant stakeholders to **jointly assess the most important ethical priorities and risks for the use case**.
Furthermore, when presenting the framework and methodology to the other projects ETAPAS collaborates with, such as for example IMPULSE, which deals with **ID management**, and popAI, which focuses on **law enforcement and policing**, the completeness and usefulness of the framework **for all sectors** was once again validated.

3. ALTAI

The Assessment List for Trustworthy Artificial Intelligence (**ALTAI**) is a practical tool that helps business and organisations to self-assess the trustworthiness of their AI systems under development.

Thanks to ALTAI, AI principles determined by the High-Level Expert Group on AI (AI HLEG), are translated into an accessible and dynamic checklist that guides developers and deployers of AI in implementing such principles in practice. ALTAI is structured to ensure that users can benefit from AI avoiding unnecessary risks by indicating a set of concrete steps for self-assessment.

ALTAI's goal is to provide a complete evaluation process to conduct a Trustworthy AI self-evaluation. Organisations can draw elements relevant to the specific AI system from ALTAI or add elements to it as they see fit, taking into consideration the sector they operate in. It helps organisations understand what Trustworthy AI is, especially focusing on the risks an AI system might generate. It should raise awareness of the potential impact of AI on society, the environment, consumers, workers, and citizens, promoting the involvement of all relevant stakeholders, and helping gaining insight on whether meaningful and appropriate solutions or processes to accomplish adherence to the requirements are already in place or need to be put in place.

That said, we will consider concrete applications of ALTAI in different sectors to understand the way the self-assessment tool has been used and how it has been received by industry players.

3.1 The Assessment List on Trustworthy AI

ALTAI was developed over two years, from June 2018 to June 2020, by the High-Level Expert Group on AI (HLEG), with a first piloting version of the assessment itself being released during the second half of 2019.

As mentioned before, adopting a trustworthy approach is key to enabling "responsible competitiveness", and ALTAI provides the foundation upon which all those who use or enter in contact with AI systems can confidently trust that their design, development, and use are lawful, ethical, and forceful. Therefore, ALTAI supports spreading responsible and sustainable AI innovation in Europe. Moreover, ALTAI also seeks to make ethics a core pillar for developing a unique approach to AI, one aimed to benefit, empower, and protect human development and the common good of society. This way, Europe and European organisations are enabled to become global leaders in cutting-edge AI worthy of both individual and collective trust.

To reach such objectives, the concept of Trustworthy AI was introduced by the AI HLEG in the Ethics Guidelines for Trustworthy Artificial Intelligence (AI), which also presented the seven key requirements on which ALTAI's self-assessment is based[2]:

1. **Human Agency and Oversight**: fundamental rights, human agency, and human oversight.
2. **Technical Robustness and Safety**: resilience to attack and security, fall back plan and general safety, accuracy, reliability, and reproducibility.
3. **Privacy and Data Governance**: respect for privacy, quality and integrity of data, access to data.
4. **Transparency**: traceability, explainability, communication.
5. **Diversity, Non-discrimination, and Fairness**: avoidance of unfair bias, accessibility, and universal design.
6. **Societal and Environmental Well-being**: sustainability and environmental friendliness, social impact, society, and democracy.
7. **Accountability**: auditability, minimization and reporting of negative impact, trade-offs, and redress.

Such requirements are embedded into the Assessment List, as they are used create a questionnaire that users must fill in to complete the self-assessment.

The questionnaire is structured following the order according to which the seven requirements have been presented, with each question then leading to different scenarios depending on the answer given, as multiple-choice questions lead to multiple different configurations of the same questionnaire, with key questions remaining unchanged. Moreover, each section of the questionnaire ends with a short self-evaluation related to the specific pillar that is being checked, which users must complete to confirm and end the questionnaire to immediately receive advice on how to implement the matter in topic in the users' AI-based solutions.

Then, upon concluding the overall ALTAI self-assessment questionnaire, users will be given a visualisation of the self-assessed level of adherence of the AI system with the seven requirements for Trustworthy AI so to identify areas for further improvement, and recommendations based on the answers to particularly significative questions.

[2] https://altai.insight-centre.org/

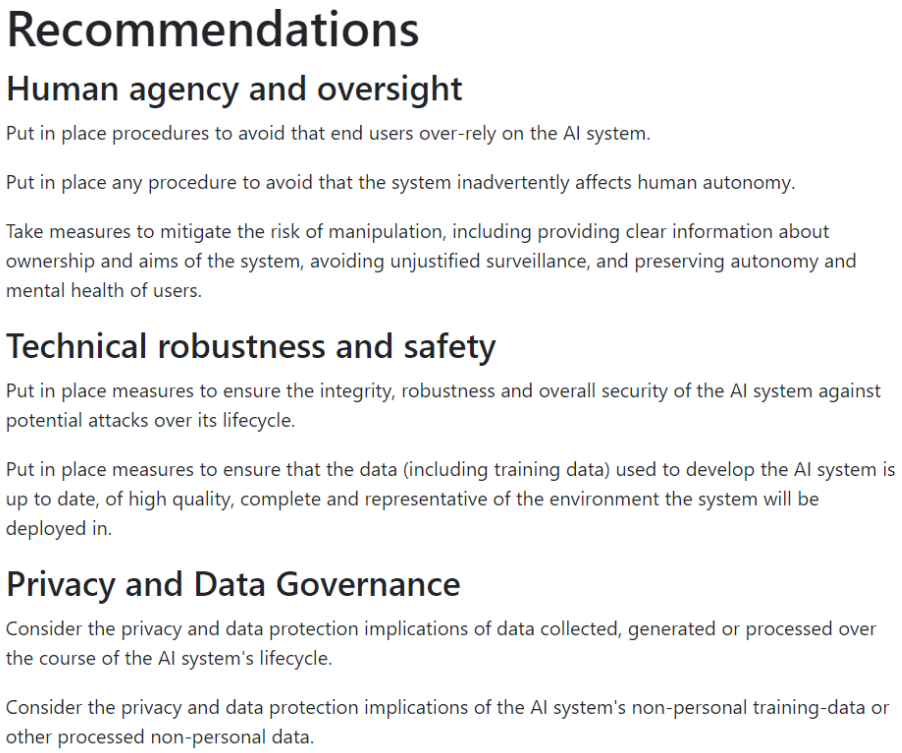

Figure 5 - Recommendations provided by ALTAI

Finally, as anticipated before, our study has considered concrete applications of ALTAI in different sectors by analysing the existing literature on this topic. To this extent, desk research has been carried out identifying 26 papers which highlight how ALTAI has been implemented to conduct self-evaluations concerning the risks using AI-based solutions generates. Therefore, from the research we led, several sectoral applications of ALTAI emerged, with the most relevant ones being represented by applying the self-assessment tool in **Commerce**, **Cybersecurity**, **Earth Observation, Energy, Education**, **Engineering, Food**, **Health**, **Legal**, **Manufacturing, Military, and Transport**[3].

3.2 Feedbacks and lessons learned from ALTAI users

Analysing the feedback of ALTAI users both from the literature on the subject and collected directly from organisations participating in projects with which ETAPAS collaborates, we have evidence of several merits and a number of limitations of the assessment tool.

Indeed, starting from the literature analysed on ALTAI concerning Education, most authors agree on the fact that ALTAI requirements can be embedded form the **very design stage** to develop AI-based dashboards and trustworthy warning systems to monitor university students' and professors' activities thus ensuring the adoption of a fair and honest behaviour. Despite that, ALTAI lack of specificity with regards to the educational context represents one of the main criticalities of the self-assessment tool, given that, as anticipated above, **only general and non-mandatory suggestions** are provided to ALTAI users.

[3] All documents are indicated in the References section

Moreover, focusing on the existing literature on the use of ALTAI in **healthcare**, ALTAI requirements can be applied so to improve the traceability of data when data drift occurs in the domain of medical imaging, help to improve setting-up Machine Learning systems in healthcare, and prevent data breaches in the field of genomics data collection. To this extent, representing a safe and well-planned solution to be applied in the healthcare sector. Again, the main limit of ALTAI is indeed represented by the already mentioned **lack of specificity**, with another criticality being represented by the fact that ALTAI does not provide any explanation for the advice it gives, an element which must be granted by EU law whenever AI is used in healthcare.

In addition, considering the literature produced on the implementation of ALTAI in the **legal** sector, if ALTAI requirements could be integrated in systems and frameworks aimed at regulating many aspects of life, from granting and respecting basic human rights, to controlling complex technologies like Quantum technologies, using the self-assessment tool would still remain suspicious, as the tool might be exploited by specific committees, with the members of such committees manipulating ALTAI to obtain false results, thus only complying with ALTAI requirements on paper with the aim to obtain and gain reputational benefits.

Further ALTAI criticalities have been also highlighted the existing literature in the field of **cybersecurity**, given that, since the pilot version of the self-assessment tool has been developed, ALTAI has limited uses in this sector. In fact, ALTAI misses adversarial attack methods against machine learning models. Indeed, monitoring systems that utilise machine-learning models should be embedded in the tool to detect inputs indicative of adversarial attacks. It is then necessary to verify whether this specific feature has been integrated with ALTAI.

Literature has also reflected on how ALTAI can be used to verify how reliable is information provided by companies to consumers with regards to their products' sustainability levels and their commercial practices. Considering the first aspect, namely sustainability, what emerges from literature is the fact that ALTAI cannot be only tool to carry such analysis, as it only provides incomplete feedbacks that do not turn into mandatory measures to be applied and respected by firms, with soft law being too weak to grant reliability and truthfulness in such sector. Moreover, analysing firms' commercial practices, commercial AI-based chatbots and algorithmic management systems have been studied. From this perspective, it is possible to observe how Trade Unions can exploit the requirement of transparency introduced by ALTAI, therefore using AI to inform users from the beginning about the fact that they are dealing with an artificial intelligence, but these guidelines are indeed set for more powerful tools than Chatbots, therefore, by applying ALTAI requirements, it has been observed how **Chatbots** perform very poorly.

Finally, despite these criticalities, authors generally affirm how adherence to ALTAI guidelines is key to secure the adoption of trustworthy AI, as the requirements embedded in the self-assessment tool are not complicated to be incorporated in the users' activities.

This said, feedback on ALTAI was also gathered from other European projects that have concretely applied ALTAI self-assessment to verify their compliance with ALTAI requirements. Analysing the feedback collected from their experience, users agree on the fact that having a common assessment framework is crucial to evaluate the projects' performance, with the common assessment approach also allowing to share the outcomes, challenges faced, and the best practices coming from the different projects. What stands out from these projects' perspective is that, by benefitting from the expertise of ALTAI authors as well as the support of

the European Commission on the ALTAI, **credibility and trust** are provided to the overall procedure, thus validating ALTAI also to the general public.

Then, the EU projects also underlined ALTAI significant limitations, especially the ones that concern the way the questionnaire is structured and its specificity in terms of application. Firstly, taking into account the questionnaire's structure, the guidelines to be followed to fill in the questionnaire are **complex and time-wasting**, with basic knowledge about ethical issues being mandatorily requested. Moreover, the Projects also highlighted that the questionnaire is built in a way that leads most of the respondents to act defensively, and, although most questions seem simple at first sight, replying to each one of them requires significant infrastructure, tools, and expertise, resources that small SMEs and Start-ups might not possess. Finally, EU projects also noted how respondents skipped many questions highlighting how a **self-assessment tool**, as ALTAI is intended to be used, cannot be enough and the results should be cross-checked by external reviewers.

4. ETAPAS – ALTAI Comparative Analysis

ETAPAS and ALTAI present assessment methodologies which are significantly different even though the goal pursued by the tools can be easily compared. Therefore, a comparative analysis can be carried out to observe these differences and then point out the strengths and weaknesses of both methodologies, with the goal of identifying the way such hindrances can be overcome and guide PAs towards an improved assessment process.

To lead this comparative analysis, we set up data-collection activities which involved different methodologies. Concerning **ETAPAS**, information on the tool was collected by consulting project deliverables **D2.1, D2.2, D.23, D3.1, D3.2,** and **D3.3**[4], while with regards to **ALTAI**, crucial information was obtained by consulting **ALTAI website and attached documentation**[5], as they describe the assessment procedures in a clear and precise way. In addition, we also collected **feedback form users**, namely from the four ETAPAS real-life use cases, the EU projects collaborating with ETAPAS and their Consortia and we analysed the existing **literature** developed on the topic and for which desk research has been carried out studying papers and publications specifically focused on the topic of Trustworthy AI and how it can be ensured using ALTAI self-assessment tool in different sectors. Therefore, secondary data has been collected to build such comparison, which is then focused on analysing the main strengths and weaknesses of the assessment procedures. Strengths and weaknesses have been identified based on the **desirable features** for a trustworthy AI assessment emerged from the users' feedback. The methodology followed for the Comparative Analysis is briefly depicted in Figure 5 below.

[4] More on the Project Deliverables can be found in the References section
[5] https://altai.insight-centre.org/

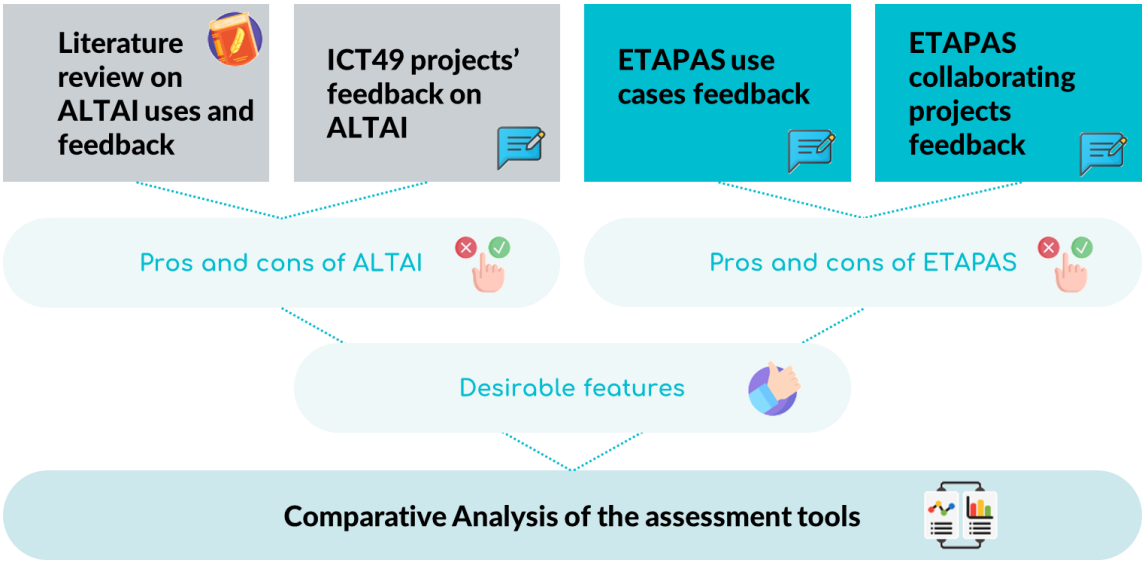

Figure 5 - Comparative analysis methodology

The following sections focus on the pros and cons of the assessment methodologies and the lessons learnt from their comparative analysis. To introduce them, the two methodologies are also briefly described below.

4.1 Pros and cons of ALTAI and ETAPAS

When comparing ALTAI and ETAPAS assessment methodologies, it is possible to consider how both assessment tools present many commonalities, such as being based on ethical principles and the check-list configuration. However, they also show many key differences, especially in the way of delivery and in the information considered.

Commonalities

Reliability and Trust

Both ALTAI and ETAPAS present strong theoretical backgrounds on which they are founded. In fact, ALTAI main strength is indeed represented by the credibility of its authors, as those who contributed to identifying the requirements upon which basing ALTAI questionnaire are part of the High-Level Expert Group on Artificial Intelligence (AI HLEG), which has worked closely with the European community of AI stakeholders through the AI Alliance in several different circumstances to define the legal requirements to be applied to the ethical use of Artificial Intelligence. Simultaneously, ETAPAS is based on several studies - led on the topic of Trustworthy AI - and analysis of best practices, including ALTAI itself, and embraces heterogenous topics which space from ethical principles to social and legal impacts. Therefore, both tools are reliable when used by end-users and lead to trustable result.

Presence of detailed guidelines

ALTAI and ETAPAS both offer a series of clear and precise guidelines that are made available to users so to make the process easier to be completed. In fact, ALTAI presents the 7 requirements mentioned before, namely **Human Agency and Oversight**, **Technical Robustness and Safety**, **Privacy and Data Governance**, **Transparency**, **Diversity, Non-discrimination, and Fairness**,

Societal and Environmental Well-being, **Accountability**, that act as guidelines that users need to follow to fill in ALTAI questionnaire, as it is structured respecting the order according to which the requirements are presented.

On the other hand, ETAPAS has a wide **Code of Conduct** that is regulates the assessment procedure and supports users to ensure their compliance with Trustworthy AI criteria and requirements.

Suggesting risk mitigation actions

As soon as the assessment procedure ends, both ALTAI and ETAPAS produce recommendations aimed at underlining the strengths and weaknesses users present considering the fields tested, while also providing immediate advice on how to cope with the risks presented by weak Trustworthy AI systems. Additionally, defensive mechanisms are also suggested so to be applied as soon as possible depending on the users' available resources.

Applicability from the design stage

As introduced when studying the existing literature on ALTAI concrete use-cases, the self-assessment tool can be directly implemented since the design stage, with little to no hindrances met when doing so. Indeed, ALTAI can be immediately embedded in the AI-based solutions since its inception. Similarly, ETAPAS can also be implemented since the design stage when developing disruptive technologies, with the early adoption of the **generic Code of Conduct for the adoption of DTs in the Public Sector** that regulates ETAPAS assessment procedure.

Comparability of results over time

After having completed the assessment, both ALTAI and ETAPAS offer users the possibility to save their results for future evaluations and comparisons over time. Indeed, results are available and can be used to observe the increased or decreased adherence to Trustworthy AI principles determined by ALTAI and ETAPAS, depending on the assessment tool users choose.

Differences

Applicability and Methodology

The first difference that can be observed considering ALTAI and ETAPAS regards the applicability of such tools. In fact, ALTAI can only be used when evaluating AI-based solutions. On the other hand, ETAPAS application range is wider, as it includes all disruptive technologies. Consequently, ETAPAS approach can be extended to a wider number of use-cases compared to ALTAI, which is instead limited to a niche of use-cases.

Then, moving to the methodology both tools envisage, ALTAI methodology moves from the aforementioned 7 requirements to complete ALTAI questionnaire whose results are immediately released. Concerning such results, ALTAI only provides recommendations which are not mandatory to be implemented by its users. As a matter of fact, ALTAI is a voluntary self-assessment tool entities can decide to undertake to verify whether the outputs of the activities they carry out, namely AI-based products and services, comply with the ethical requirements discussed above. Consequently, this lack of compulsory actions to be taken after ALTAI feedback

has been received is considered as a crucial weakness, with such criticality also being greatly underlined by the literature on ALTAI.

On the contrary, ETAPAS is based upon an extensive risk mapping procedure, a deep analysis of the European legal framework, in addition to the 10 ethical principles of the ETAPAS Code of Conduct, all specific to Public Administration, and the assessment's indicators are mapped on both ethical principles (10), risks (34) and risk categories (8), ensuring wide coverage. Moreover, ETAPAS methodology allows users to recognise the most significant risk categories identified for specific use-cases, thus assigning a higher weight to these risks. Then, after having completed ETAPAS assessment leads to a **more objective and detailed analysis** which is deployed as the starting point for further support received by users so to implement the suggested solutions.

Easy-to-use online tool

Despite ALTAI and ETAPAS both being online tools, one main difference can be observed when taking into consideration this dimension.

In fact, ALTAI is publicly available, and anyone can complete the self-assessment procedure by filling in the questionnaire presented on ALTAI website at this link.

On the other hand, ETAPAS assessment can be accessed only through a private platform, developed overtime and yet to be completely finalised, which will then offer additional and tailor-made services to support users during the evaluation.

Possibility of tailoring to the implementation stage/ maturity of the AI solution

The ETAPAS approach is designed so to embrace the entire life cycle of the AI solution, starting from the design stage, continuing into the implementation phase and beyond, and being tailored for each maturity stage of the AI solution in question. Indeed, ETAPAS approach allows users to receive customised solutions according to the solution's maturity level.

On the other hand, ALTAI only presents a generic questionnaire that cannot be changed in its overall structure, which then results to be fixed and non-suitable to the users' needs at the different maturity levels of the AI solution.

Clear scoring system

Considering ALTAI and ETAPAS scoring systems, the latter appears to be more complete compared to the one ALTAI uses.

In fact, ETAPAS adopts a weight-based scoring model which allows to distinguish and recognise the different risks users run by taking into account their gravity, assigning them different weights, as illustrated before. This way, the most relevant risks can be tackled first, without incurring in time-wasting operations which then might worsen more serious menaces.

ALTAI scoring system, instead, is simply represented by a matrix where all requirements are assigned a score ranging from 0 to 5 that indicates how strong the user is in that specific area. Indeed, such scores are only used as the basis on which then providing general recommendations which act as suggestions on how to improve in these areas according to the answers previously recorded when completing ALTAI questionnaire.

Requiring little time

Considering the time needed to complete the whole assessment process, ETAPAS requires greater effort from its users who often need to be supported, especially at the beginning and, even if to a lesser extent, along the whole assessment path. On the other hand, ALTAI questionnaire and self-assessment procedure can be completed faster, with few questions requiring open-ended answers, given that most of them are multiple-choice questions.

Methodology for stakeholder involvement

Regarding stakeholder involvement, significant differences between ALTAI and ETAPAS need to be acknowledged. Indeed, according to the literature we collected on ALTAI, it is suggested that ALTAI should involve not only EU-based firms, but also EU citizens, thus empowering and allowing them to generate a significant impact on the way companies operate in this environment. On the contrary, citizens are left behind with only European companies being involved in the process to shape the ethical use of AI solutions.

From this perspective, ETAPAS approach, instead, includes stakeholder engagement techniques since the very inception stage, involving citizens and asking for their feedback and perception of the disruptive tool.

Automatic monitoring of results over time

As mentioned before, the ETAPAS approach embraces the entire life cycle of an AI solution, starting from the design stage, continuing into the implementation phase and beyond, allowing automatic monitoring of computational indicators and suggesting at least three different questionnaires. ALTAI, instead, does not allow this constant monitoring of results, as a new questionnaire needs to be filled in from scratch after a previous form has already been completed.

Concluding, Table 1 below summarises the results of the comparative analysis between the two trustworthy AI assessment methodologies.

Desirable features	**ETAPAS**	**ALTAI**
Reliability and Trust	✅	✅
Applicable from the design stage	✅	✅
Possibility of tailoring to the implementation stage/ maturity of the AI solution	✅	❌
Presence of detailed guidelines	✅	✅
Clear scoring system	✅	❌
Possibility of tailoring to the context and characteristics of the AI application	✅	❌
Easy-to-use online tool	◐	✅
Requiring little time	❌	◐

Methodology for stakeholder involvement	✓	✗
Automatic monitoring of results over time	✓	✗
Results comparable over time	✓	✓
Suggesting risk mitigation actions	✓	✓

Table 1 - ETAPAS - ALTAI comparison

4.2 Lessons learned from the comparative analysis

To finalise this reflection on the differences between ALTAI and ETAPAS, it is possible to see how ETAPAS assessment tool **covers a wider range of aspects** compared to ALTAI given the specific nature of the tools themselves. To this extent, ALTAI is specifically designed to be applied to AI-based solutions, while the ETAPAS assessment tool moved from the basis provided by ALTAI, then implementing additional coverage given to almost every disruptive technology. Consequently, the ETAPAS assessment tool results to be **more complex** compared to ALTAI, with an increased **precision and reliability** gained in the field of assessing the compliance with ethical requirements to set up Trustworthy AI solutions. In addition, providing concrete and detailed recommendations makes it **easier for ETAPAS users to develop strategies and mitigation actions** aimed at preventing the risks related to the unethical use of AI. Therefore, compared to ALTAI, ETAPAS offers an assessment tool that produces more concrete consequences as where ALTAI provides general recommendations, ETAPAS allows to detect specific criticalities and to correct them following ETAPAS own recommendations. It can also be said that the ETAPAS methodology is **more targeted at the public sector** and those companies that are likely to have high social, ethical, and legal impacts. On the other hand, ALTAI is more suitable for the initial design phase of AI solutions and for applications whose ethical analysis is more straightforward.

5. Policy Recommendations

To support the ethical adoption of Disruptive Technologies in PAs avoiding the risks embedding AI presents, preventive measures should be ensured so to be used since the solution's very inception. To this extent, it would be necessary for such measures to be applied to all disruptive technologies, shall allow to run an ethical, social, and legal evaluation from the very design phase of the AI-based application, involve all stakeholders getting their contribution to suggest criticalities and eventual improvements, offer trustable and reliable advice that allow to quickly put in place risk-prevention actions, and that lead to results that can be monitored over time. The following paragraphs will delve into the details of all the aforementioned features to provide meaningful recommendations.

5.1 Apply the assessment process to all Disruptive Technologies and tailor it to the different sectors

Given their nature, disruptive technologies - here also including AI – shall be monitored to prevent and tackle the risks stemming from them. Therefore, limiting the application of assessment tools

aimed at reaching the goal mentioned above to a single category of application, namely AI-based solutions, would only allow a niche of users to run these monitoring activities, thus increasing the probability that the negative consequences generated by such technologies could greatly affect unaware users. A symmetrical problem is that such an assessment must be applied to different Disruptive Technologies in different sectors, which may have different peculiar characteristics. Therefore, a tailoring process for different sectors may help the adoption of the assessment by public administrations.

5.2 Start the ethical, social, and legal assessment from the design phase of the application and throughout the application's entire lifecycle

AI-based solutions tend to rapidly evolve, undergoing frequent changes that might significantly impact and modify their nature. Consequently, the risks related to such applications must be monitored since the design phase of these solutions. To this extent, developing an assessment tool which can be employed in the initial stages of the application's life cycle, would allow to have a preliminary overview of the risks which characterise that solution, so to prevent them before they turn into more serious menaces. In addition, by starting the assessment this early, risks may also be estimated beforehand, thus developing defensive measures to be applied so to adhere to the principles that determine a trustworthy use of AI. This said, monitoring AI-based solutions since the design phase shall be part of a wider monitoring plan aimed at conducting a systematic ethical, social, and legal assessment covering the application's entire lifecycle. By doing so, the unpredicted changes mentioned at the beginning of the paragraph would be detected in an iterative way, thus resulting in a timelier response to unplanned variations in the application's nature.

5.3 Involve all relevant stakeholders in the assessment process, including users and final users

Stakeholders' involvement is crucial to develop successful assessment tools. Indeed, by getting the contribution of final users of AI-based applications, direct feedback will be gathered on the impact AI itself has on their life and economic activities, with such feedback potentially acting as the basis on which making further progress and improving the assessment tools needed to evaluate AI. Consequently, ensuring a state-of-the-art level of compliance with trustworthy AI principles would be in their interest, given that, otherwise, final users would be severely damaged. Therefore - as mentioned above - exploiting the inputs coming from end-users first and stakeholders in general could only lead to more precise and reliable evaluation tools to be applied when considering the use of AI and disruptive technologies.

5.4 Ensure reliable results

Another fundamental aspect of an evaluation tool is represented by the results that the assessment itself produces. From this perspective, when an analysis to check whether an AI-based solution responds to the standards required to be defined as a trustworthy application is carried out, the results that are obtained by those who run the evaluation must present a solid theoretical background which shall also be combined with a fact-based nature. Indeed, such nature can be only obtained through fact-based questions that – during the evaluation – will allow to understand which actions have been carried out in the development of the AI solution, thus ensuring objective answers. Consequently, the recommendations generated after the assessment would not only consider the most relevant theories developed in the field of preventing and fighting AI-related risks but would also reflect on the actions taken by the entity

that is running the analysis, so to generate a tailor-made solution that can almost perfectly fit the specific case in question. To reach such a result, it is necessary for the structure of the assessment tool to be based on both theoretical principles and practical activities that users are supposed to undertake to complete the evaluation.

5.5 Monitor results over time

As mentioned in the paragraphs above, assessments aimed to verify the trustworthy use of AI solutions and applications should be performed according to a systematic monitoring plan detailed by final users of such technologies. To this extent, as multiple assessments will be conducted over time, users shall be able to record the results obtained for each evaluation, so to compare them to the outputs generated by previous assessments and analyse the existing differences to understand whether improvements were made, or fallouts took place, with the final goal of making the AI solution trustworthy. To this extent, higher comparability of results across time would be gained by adopting a scoring system that presents and maintains fixed elements despite the changes the AI solution might undergo. Such elements, together with the tailor-made components mentioned in the previous paragraph, can be used to make comparisons among different evaluations ran in different moments, with fully objectivity of judgement being ensured in case perfectly comparable results coming from different assessments are recorded.

6. References

A. Serban, K. van der Blom, H. Hoos and J. Visser, "Practices for Engineering Trustworthy Machine Learning Applications," 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN), 2021, pp. 97-100, doi: 10.1109/WAIN52551.2021.00021.

Amram D, Cignoni A, Banfi T et al. From P4 medicine to P5 medicine: transitional times for a more human-centric approach to AI-based tools for hospitals of tomorrow [version 1; peer review: 2 approved]. Open Res Europe 2022, 2:33 (https://doi.org/10.12688/openreseurope.14524.1)

Applying Ethical AI Frameworks in practice: Evaluating conversational AI chatbot solutions. S Atkins, I Badrie, S van Otterloo – 2021

Baneres D, Guerrero-Roldán AE, Rodríguez-González ME, Karadeniz A. A Predictive Analytics Infrastructure to Support a Trustworthy Early Warning System. Applied Sciences. 2021; 11(13):5781. https://doi.org/10.3390/app11135781

Bouchon-Meunier, B., "Ethics, Diversity and Consciousness in AI [President's Message]," in IEEE Computational Intelligence Magazine, vol. 16, no. 3, pp. 3-4, Aug. 2021, doi: 10.1109/MCI.2021.3084388.

D. Scaradozzi, L. Screpanti and L. Cesaretti, "Machine Learning for modelling and identification of Educational Robotics activities," 2021 29th Mediterranean Conference on Control and Automation (MED), 2021, pp. 753-758, doi: 10.1109/MED51440.2021.9480309.

Fernandez Llorca, D. and Gomez Gutierrez, E., Trustworthy Autonomous Vehicles, EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-46055-8, doi:10.2760/120385, JRC127051

Fröding, Carvalho, Mureddu, Hansson, Zamichos, and Kougioumtzidou. “Deliverable 2.1, Literature review on ethical, legal and social aspects of disruptive technologies (DTs)”, June 2021.

Fröhlich Holger, Bontridder Noémi, Petrovska-Delacréta Dijana, Glaab Enrico, Kluge Felix, Yacoubi Mounim El, Marín Valero Mayca, Corvol Jean-Christophe, Eskofier Bjoern, Van Gyseghem Jean-Marc, Lehericy Stepháne, Winkler Jürgen, Klucken Jochen. Leveraging the Potential of Digital Technology for Better Individualized Treatment of Parkinson's Disease, Frontiers in Neurology, Volume 13, 2022, https://www.frontiersin.org/article/10.3389/fneur.2022.788427, doi: 10.3389/fneur.2022.788427, ISSN: 1664-2295

FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging. Karim Lekadir, Richard Osuala, Catherine Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Aussó, Leonor Cerdá Alberich, Kostas Marias, Manolis Tsiknakis, Sara Colantonio, Nickolas Papanikolaou, Zohaib Salahuddin, Henry C Woodruff, Philippe Lambin, Luis Martí-Bonmatí

Gottardo, R. (2021) 'Building Global Algorithmic Accountability Regimes: A Future-focused Human Rights Agenda Beyond Measurement', Peace Human Rights Governance, 5(1), 65-96.

Hansson, Fröding. "Deliverable 2.2, Code of Conduct for DTs," April 2021.

Jordi Albo-Canals, Denise Amram, Katharina Kaesling, Juan Martinez Otero, Ruggero G. Pensa, and Olga Sans-Cope. 2021. Children's Rights in Online Environments with Social Robots: The use case study of CORP: A Collaborative Online Robotics Platform. In Proceedings of Child-Robot Interaction Child's Fundamental Rights Workshop in conjunction withthe ACM International Conference of Human-Robot Interaction (HRI2021) - (HRI '21). ACM, New York, NY, USA, 5 pages.

Kindylidi I, Cabral TS. Sustainability of AI: The Case of Provision of Information to Consumers. Sustainability. 2021; 13(21):12064. https://doi.org/10.3390/su132112064

Kop, Mauritz, Establishing a Legal-Ethical Framework for Quantum Technology (March 2, 2021). Yale Law School, Yale Journal of Law & Technology (YJoLT), The Record, March 30, 2021, https://yjolt.org/blog/establishing-legal-ethical-framework-quantum-technology, Available at SSRN: https://ssrn.com/abstract=3814422

Larasati, Retno; De Liddo, Anna and Motta, Enrico (2021). AI Healthcare System Interface: Explanation Design for Non-Expert User Trust. In: ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops (Glowacka, Dorota and Krishnamurthy, Vinayak eds.), CEUR Workshop Proceedings, 2903.

La Rosa, Pellegrino, Brunelleschi, Mancini, Rutta, Carmeno, Tsourma, and Kougioumtzidou. "Deliverable 3.2, Measurement Guidelines," October 2021.

Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Frontiers in Human Dynamics, 3, [673104]. https://doi.org/10.17863/CAM.73264

Levina, O., Towards Implementation of Ethical Issues into the Recommender Systems Design, Technische Hochschule Brandenburg, University of Applied Science, Brandenburg, Germany.

M. Borg, R. Jabangwe, S. Åberg, A. Ekblom, L. Hedlund and A. Lidfeldt, "Test Automation with Grad-CAM Heatmaps - A Future Pipe Segment in MLOps for Vision AI?" 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2021, pp. 175-181, doi: 10.1109/ICSTW52544.2021.00039

M. Borg et al., "Exploring the Assessment List for Trustworthy AI in the Context of Advanced Driver-Assistance Systems," 2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics), 2021, pp. 5-12, doi: 10.1109/SEthics52569.2021.00009.

Mancini, Hansson, Carmeno. “Deliverable 2.3, Report on potential risks of DTs in public administration,” June 2021.

Mureddu, Moise, Mancini, Galasso, Carmeno, and Rutta. “Deliverable 3.3, Validation method and guidelines”, October 2021.

Nowik, P. (2022), “New challenges for trade unions in the face of algorithmic management in the work environment”, Studies on Labour Law and Social Policy.

Oestreich M, Chen D, Schultze JL, Fritz M, Becker M. Privacy considerations for sharing genomics data. EXCLI J. 2021; 20:1243-1260. Published 2021 Jul 16. doi:10.17179/excli2021-4002

Pellegrino, Galasso, Brunelleschi, Mancini, Rutta, Carmeno, Tsourma, and Kougioumtzidou. “Deliverable 3.1, Responsible DT indicators and metrics”, October 2021.

Pupillo, Lorenzo & Ferreira, Afonso & Fantin, Stefano, 2020. "Artificial Intelligence and Cybersecurity - Task Force Evaluation of the HLEG Trustworthy AI Assessment List (Pilot Version)," CEPS Papers 26204, Centre for European Policy Studies.

Responsible and Ethical Military AI, Allies and Allied Perspectives, CSET Issue Brief, 2021

Rowena Rodrigues, Legal and human rights issues of AI: Gaps, challenges and vulnerabilities, Journal of Responsible Technology, Volume 4, 2020, 100005, ISSN 2666-6596, https://doi.org/10.1016/j.jrt.2020.100005

Smuha, Nathalie A., Trustworthy Artificial Intelligence in Education: Pitfalls and Pathways (December 2020). Available at SSRN: https://ssrn.com/abstract=3742421 or http://dx.doi.org/10.2139/ssrn.3742421

Van Dijk, N., Casiraghi, N., and Gutwirth, S., The ‘Ethification’ of ICT Governance. Artificial Intelligence and Data Protection in the European Union, Computer Law & Security Review, Volume 43, 2021, 105597, ISSN 0267-3649, https://doi.org/10.1016/j.clsr.2021.105597

Vetter, D., Westerlund, M., ... Kararigas, G. (2021). On Assessing Trustworthy AI in Healthcare: Machine