

AI-CODE

AI services for COntinuous trust in emerging Digital Enviroments

Deliverable 9.5

Ethical Design Requirements

Grant Agreement no.

101135437

Project Title

AI-CODE - AI services for COntinuous trust in emerging Digital Environments

Contractual Due date

30 November 2025 (M24)

Actual submission date:

26 November 2025

Lead beneficiary:

CEPS



Deliverable number	9.5
Deliverable Title:	Ethical Design Requirements
Description:	This deliverable describes the ethics-by-design framework applied in this project. It includes an overall description of the framework, situating this work within a larger body of applied research. Beyond the identification of ethical design principles, this project conducted a mapping of ethical issues and received valuable contributions from the Ethics Advisory Board (EAB) which is described in this deliverable. Finally, the deliverable includes a preliminary assessment on the progress of implementing the chosen ethical principles.
Type/Nature	☒ R - Document, report ☐ DEM - <i>Demonstrator, pilot, prototype</i> ☐ DEC - <i>Websites, patent filings, videos, etc.</i> ☐ OTHER ☐ ETHICS – <i>Ethics requirement</i> ☐ ORDP – <i>Open Research Data Pilot</i>
Dissemination Level	☒ Public – fully open ☐ Sensitive — limited under the conditions of the Grant Agreement ☐ EU classified — RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444. For items classified under other rules (e.g. national or international organization), please select the equivalent EU classification level.
Work Package number	WP9
Work Package leader	DST
Author(s) <i>in alphabetical order</i>	Aires, Susana (CEPS) Armengol, Abril (CEPS) Griffith, Jacob (CEPS) Gürtler, Paula (CEPS) Renda, Andrea (CEPS)
Internal Reviewers <i>in alphabetical order</i>	Albrech, Bahia (DW) Boi, Siliva (NAL)

AUTHORS

Name	Partner	E-Mail
Paula Gürtler	CEPS	paula.gurtler@ceps.eu
Susana Aires	CEPS	susana.airesgomes@ceps.eu
Andrea Renda	CEPS	andrea.renda@ceps.eu
Jacob Griffith	CEPS	jacob.griffith@ceps.eu
Abril Armengol	CEPS	--

HISTORY OF CHANGES

Version	Date	Changes
1.0	20/10/2025	First draft
1.1	7/11/2025	Complete version for internal peer review
2.0	25/11/2025	Final version for submission

TABLE OF CONTENT

Executive summary	1
1. Introduction: Ethics-by-design.....	2
1.1 What is ethics by design?	3
1.2 Why do ethics by design?	4
1.3 Addressing shortcomings of EbD in AI-CODE	5
1.4 Deliverable outline	6
2. Ethics-by-design in the project	7
2.1 AI-CODE Services	7
2.1.1 Data-driven services.....	7
2.1.2 User-driven services.....	8
2.2 Mapping areas of concern.....	8
2.2.1 Exploratory literature review to identify general ethical issues.....	9
2.2.2 Visual Mapping of ethical issues	10
2.3 The Ethics Advisory Board	11
2.3.1 The first annual EAB meeting (M12).....	12
2.3.2 Service-specific feedback from EAB (M12).....	12
2.3.3 Service 1: Disinformation detection in the next-generation social media (T4.1/5.1) .	13
2.3.4 Service 2: Trustability and credibility assessment (T4.2/5.2)	14
2.3.5 Service 3: Media Assessment and Annotation Management (MAAM) (T4.3/5.3)	14
2.3.6 Service 4: Generative AI Interactive Coaching and Simulation & Counter Narrative Recommendation Service (T6.1/7.1)	15
2.3.7 Service 5: Transparency Services for AI-CODE Services (T6.2/7.2).....	16
2.3.8 Service 6: Personal Companion for understanding disinformation for media professionals (T6.3).....	16
2.4 Ethical design requirements.....	17
2.4.1 The AI-HLEG Ethics Guidelines for Trustworthy AI	17
2.4.2 From Guidelines to Design Principles	18
2.5 Evaluation with Assessment List for Trustworthy AI (ALTAI)	18
2.5.1 Method in the project	19
3. Preliminary results on implementing ethical design requirements	20
3.1 Service 1 (T4.1): Preliminary ALTAI results	21
3.2 Service 2 (T4.2): Preliminary ALTAI results	22
3.3 Service 3 (T4.3): Preliminary ALTAI results	23
3.4 Service 4a (T6.1): Preliminary ALTAI results.....	24
3.5 Service 4b (T6.1): Preliminary ALTAI results	25
3.6 Service 5 (T6.2): Preliminary ALTAI results	26
3.7 Service 6 (T6.3): Preliminary ALTAI results	28
4. Methodological Reflections.....	29

4.1	Success of EbD in AI-CODE	29
4.2	Challenges of EbD implementation	30
4.3	ALTAI: Limitations and improvements	31
5.	Conclusion.....	32
6.	References	34
Annex I – Mapping of ethical areas of concern		37
	Mapping for T4.2 Trustability and credibility assessment	37
	Mapping for T4.3 (MAAM)	38
	Mapping for T6.1 Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals.....	38
	Mapping for T6.2 Transparency Services for AI-CODE.....	39
	Mapping for T6.3 Personal Companion for understanding disinformation for media professionals	39
Annex II – Measures based on EAB Feedback		40
	T4.1 Disinformation Detection in the next generation social media	40
	T4.2 Trustability and Credibility Assessment	42
	T4.3 Media Annotation and Assessment Management (MAAM).....	43
	T6.1 Generative AI Interactive Coaching and Simulation & Counter Narrative Recommendation Service	44
	T6.2 Transparency Service for AI model cards.....	45
	T6.3 Personal Companion	47
Annex III – Ethical Design requirements		49
Annex IV – ALTAI Recommendations.....		51
	T4.1 Recommendations based on M18 ALTAI	51
	T4.2 Recommendations based on M18 ALTAI	53
	T4.3 Recommendations based on M18 ALTAI	55
	T6.1A Recommendations based on M18 ALTAI	56
	T6.1B Recommendations based on M18 ALTAI	57
	T6.2 Recommendations based on M18 ALTAI	58
	T6.3 Recommendations based on M18 ALTAI	60

TABLE OF FIGURES

Figure 1. Diagram of Ethics Approach in AI-CODE	6
Figure 2. Mapping general ethical areas of concern	11
Figure 3. Spider graph T4.1 ALTAI	21
Figure 4. Spider graph T4.2 ALTAI	22
Figure 5. Spider graph T4.3 ALTAI	24
Figure 6. Spider graph T6.1A ALTAI	25
Figure 7. Spider graph T6.1B ALTAI	26
Figure 8. Spider graph T6.2 ALTAI	27
Figure 9. Spider graph T6.3 ALTAI	28
Figure 10. Mapping of ethical concerns for T4.2	37
Figure 11. Mapping of ethical concerns for T4.3	38
Figure 12. Mapping of ethical concerns for T6.1	38
Figure 13. Mapping of ethical concerns for T6.2	39
Figure 14. Mapping of ethical concerns for T6.3	39

TABLE OF TABLES

Table 1. Summary of EAB feedback	12
Table 2. EAB recommendations for T4.1	41
Table 3. EAB recommendations for T4.2	42
Table 4. EAB recommendations for T4.3	43
Table 5. EAB recommendations for T6.1	45
Table 6. EAB recommendations for T6.2	46
Table 7. EAB recommendations for T6.3	47
Table 8. List of ethical design requirements	49
Table 9. ALTAI Recommendations T4.1 (M18)	51
Table 10. ALTAI Recommendations T4.2 (M18)	53
Table 11. ALTAI Recommendations T4.3 (M18)	55
Table 12. ALTAI Recommendations T6.1A (M18)	56
Table 13. ALTAI Recommendations T6.1B (M18)	57
Table 14. ALTAI Recommendations T6.2 (M18)	58
Table 15. ALTAI Recommendations T6.3 (M18)	60

Executive summary

This report – *D9.5 Ethics Design Requirement* – describes the ethical approach of the AI-CODE project. It is part of the broader research activities on ethics pursued during the first two years of the project under T9.3 and will ground the deliverable D9.6 Ethics Handbook, due in M36.

As described in section 1, we have applied ethics-by-Design (EbD). EbD is a framework for developing technology that systematically incorporates ethical considerations at every stage of the process. Its starting point is the recognition that ethical challenges arise during the conception and development of technologies, not just during their deployment. EbD is about raising awareness of and questioning the various ethical implications of different design choices as they are made while proactively addressing ethical issues. EbD can also draw on ethics-as-a-service approaches, including ethics advisory boards and the co-creation of ethical frameworks. The practical implementation is further supported through the use of tools that translate abstract principles into concrete criteria, such as the *assessment list for trustworthy AI* (ALTAI). EbD typically draws on common ethical principles. It is sometimes criticised for lacking domain specific considerations. In AI-CODE, we have combined a number of different measures to implement EbD, which are described in section 2.

First, we conducted a mapping of areas of concern based on a brief exploratory literature review, critical thinking, and collective brainstorming involving all consortium members (section 2.1). The emphasis was on collecting service-specific issues that can guide the design of the services. The collection of service-specific ethical issues continued with the project's Ethics Advisory Board (EAB) (section 2.2.). During the first annual meeting in month 12 (M12) of the project, the EAB provided feedback and comments on each service, pointing out specific issues that they anticipated. The discussion outcomes were reviewed and addressed with service developers in one-on-one meetings and are presented in section 2.2.2 of this report.

Section 2.3 outlines the core of the EbD approach by specifying the ethical principles that have been systematically incorporated. The *Ethics Guidelines for Trustworthy AI* by the European High-level expert group on AI (AI-HLEG) was chosen as the ethical framework. The AI-HLEG defines trustworthy AI as lawful, ethical, and robust AI systems. The approach was chosen due to its practical focus, wide-ranging applicability, significant impact on AI ethics, and consensus among experts. The AI-HLEG further specifies seven ethical requirements, which were translated into ethical design requirements for the service development (section 2.3.2). Another advantage of the *Ethics Guidelines for Trustworthy AI* is the availability of an accompanying assessment tool, i.e., the ALTAI. Our use of ALTAI is described in section 2.4 and the results of the preliminary evaluation of the services in M18 are presented in section 3. The ALTAI evaluation will be repeated in M36 to review the progress we have made towards the implementation of EbD.

Finally, this report includes a reflection on the methodological approach chosen for AI-CODE in section 4. The ethical activities have been particularly successful in raising awareness on ethical issues across the entire consortium. The service developers have faced the challenge of considering a wide range of factors beyond progressing technically. Section 4.3 specifically reflects on the limitations of the ALTAI tool for assessing the ethics of generative AI, as ALTAI was designed before this technological breakthrough. The report concludes with a short outlook for the upcoming ethical activities in M24-M36 on AI-CODE, which will be reported in the *D9.6 Ethics Handbook* due in M36.

1. Introduction: Ethics-by-design

This report introduces the approach to ethics chosen in the AI-CODE project to ensure that the six artificial intelligence (AI)-enabled services at the heart of the project are developed ethically. With the wider adoption of AI, the need to account for ethical issues – such as bias, responsibility gaps – have become apparent (e.g. Coeckelbergh, 2020). However, following ethical requirements is not only exemplary, it is also a strategy to build users’ trust in the systems and promote the adoption of these systems (Choudhury & Shamszare, 2023). Trust is especially relevant in the fields of media and journalism, where AI-CODE operates. Trust is fundamental to professional legitimacy and the important role that journalists play in democracy as providers of reliable, independent information (Deuze, 2005). Therefore, it is a priority for AI-CODE to follow good ethical practices in developing the services for media professionals.

There are many factors to consider when determining what constitutes good ethical practices. Ethics is the branch of philosophy engaged with the questions of what ought to be considered good and bad, or right and wrong. A number of normative theories, such as utilitarianism, deontology or virtue ethics can be employed to understand what is right and wrong, and there is no disciplinary consensus on which theory to employ (e.g. Kymlicka, 2002; Tännsjö, 2013).

Given the range of ethical issues that AI systems pose as they are being developed and deployed, these theories have been employed to understand what makes AI practices right or wrong. Overall, however, a principles-based approach has become the dominant approach in AI-ethics. A principles based approach (common in biomedical ethics (Beauchamp & Childress, 2009)) describes a number of value-based principles; when these are respected in the decision under observation, this decision can be deemed ethical. However, principles, such as beneficence or respect for human autonomy require definition and specification to be useful for application (Beauchamp, 2003). This contingency on specification allows for the application of already established ethical principles – such as fairness and respect for human autonomy – to be specified and applied to the context of AI.

In AI-CODE we follow the principles-based approach to ethics, as well. We rely on the ethical framework of the *Ethics Guidelines for trustworthy AI* (European Commission, 2019) by the European High Level Expert Group on AI (AI-HLEG) for selecting and specifying appropriate ethical principles. They define trustworthy AI as ethical, lawful, and robust, and provide seven requirements to determine ethical trustworthy AI: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination and fairness, 6) environmental and societal wellbeing, and 7) accountability. These principles have been used to implement an ethics-by-design approach for the AI-CODE services.

We first present ethics-by-design (EbD) as the core approach driving the ethical dimension of AI-CODE (section 1.1.) and further highlight how EbD relates to other ethics practices, namely value-sensitive design (VSD), the assessment of ethical, legal, and social aspects (ELSA) of technology, and ethics-as-a-service. In section 1.2 we explain the rationale for adopting EbD. As a practice that aims to integrate an assessment of ethical issues and implement measures to address these across all stages of technical development, EbD can effectively promote the operationalisation of AI ethics principles, including the AI HLEG’s *Ethics Guidelines for Trustworthy AI* as in this project. We single out in particular the value of pursuing ELSA and explain the rationale for integrating and applying the two in the context of AI-CODE. The final subsection of the introduction (section 1.3) provides an outline of the full deliverable.

1.1 What is ethics by design?

Ethics-by-design (EbD) is an approach to developing technology that “systematically and comprehensively” incorporates ethical considerations across all stages of technology design (Brain and Daiwon, 2023; Juhasz et al., 2025). It builds on the premise that technologies are not neutral, but sociotechnical in nature. This means that they are imagined, conceived, and developed within particular contexts and communities that actively shape the form of technical objects, how they act in the world and the social relations they mediate (e.g. Johnson and Verdicchio, 2025). EbD is not simply an isolated retrospective exercise aimed at aligning technologies with ethical requirements, as this often becomes challenging after the technologies have already been developed. Rather, the approach aims to proactively raise awareness of and question the various ethical implications of different design choices as they are made, fostering responsible practices throughout the entire technological life cycle.

While EbD precedes the advent of contemporary AI systems, with scholars pointing to the sociotechnical nature of software as early as the mid-1990s (Nissenbaum, 1996), the growing penetration of AI systems across all domains of life, their significant degree of autonomy and functional opacity, as well as the growing risks and frequency in AI-related incidents have made the complexity of our sociotechnical reality apparent. This has led to the revival of ethics as a core element of technology development. Examples include the proliferation of ethics guidelines and normative principles produced by different policy bodies and actors in both industry and academia (e.g., the OECD AI Principles and the AI HLEG’s Ethics Guidelines for Trustworthy AI; see Corrêa et al., 2023 for a systematic review), the popularity of AI alignment discussions (Gabriel, 2020) and different approaches to integrating ethical values into technology design, including EbD and value sensitive design (VSD).

As a practice for developing ethical AI systems, EbD builds on a core of ethical requirements common across guidelines and principles, including privacy, transparency, fairness, human agency, human dignity, trust, accountability and oversight, and wellbeing, among others. In contrast with VSD, the goal of EbD is not to actively mainstream particular ethical values from a user base into technology design (Van den Hoven, 2007). Instead, it focusses on assessing whether technical design practices ethically align with commonly accepted principles. These commonly agreed upon principles can vary depending on the domain of technological application and the technology in question. For the AI-CODE project – as we discuss in section 2.3 in detail – we have chosen the seven ethical principles defined by the High-Level Expert Group on AI (AI-HLEG) in the *Ethics Guidelines for Trustworthy AI* (European Commission, 2019), which are characterised by consensus among experts.

While implementing ethical principles in AI development is seen as a way to reduce the risks associated with these technologies, there are significant barriers to effectively putting them into practice. Most notably, scholars have identified a significant gap between theory and practice. AI practitioners often struggle to translate AI principles into their everyday technology development work, pointing to issues such as the high level of abstraction of ethics guidelines, lack of resources, unclear delineation of roles and responsibility in ethical assessment and implementation across the development pipeline, among others (Morley et al., 2023). This implies that EbD requires interdisciplinarity and a commitment to continuously collaborate on translating abstract principles into specific sociotechnical contexts. One aspect of EbD’s mission is to mitigate implementation challenges through providing more hands-on guidance to practitioners.

For example, “ethics-as-a-service” proposes a model of distributed responsibility to address the

challenges of embedding ethics across the AI lifecycle. In this model, multiple agents, both internal and external to the technology development organisation, co-design context-specific ethical codes of practice that enable the critical evaluation and auditing of AI development practices (Morley et al., 2021). By distributing responsibility across different actors, including external auditors and ethics advisory boards, ethics-as-a-service aims to mitigate some of the core challenges in implementing ethics procedures. Consequently, it ensures that these ethics procedures remain relevant by appropriately balancing strictness and flexibility, centralisation and delegation (Morley et al., 2021). These lessons are reflected in AI-CODE's approach to ethics which designates responsibility for providing guidance and support to CEPS and the external Ethics Advisory Board (EAB) which meets on an annual basis (see section 2.2.).

To further address the challenge of EbD, several scholars and stakeholders have put forward different methods and “translational tools” to mainstream and render AI ethics principles actionable (Morley et al., 2023). These translational tools can also include questionnaires designed to promote a dialogue about the ethical implications of specific design practices. For example, the *Assessment List for Trustworthy AI* (ALTAI), elaborated by the European Commission's AI-HLEG, comprises a range of critical questions to be addressed throughout the entire AI lifecycle (including post-deployment) aimed at ensuring AI is trustworthy, i.e. it is lawful, ethical and robust (European Commission, 2020). ALTAI strives to translate the *Ethics Guidelines for Trustworthy AI* into practical guidance and implementable procedures. This allows for an easier EbD approach to AI as promoted by the European Commission (2021). The EbD guidance by the European Commission sets out steps for each stage of technology development, namely, 1) specifying objectives of the AI system, 2) required resources, 3) designing a high-level architecture, 4) collecting and pre-processing data, 5) full system development, and 6) testing and evaluation (European Commission, 2021). A checklist for assessing ethical risks and outlining the actions taken to mitigate them is also provided. Section 2.4 of this report explains how ALTAI was utilised in AI-CODE to translate abstract ethical principles into concrete, actionable measures.

1.2 Why do ethics by design?

Building on the overview of EbD presented in the previous section, it is clear that implementing EbD is essential for responsible technology development as a whole, and is particularly crucial for ensuring trustworthy AI (Brey and Dainow, 2023; European Commission, 2021). Incorporating ethical principles throughout the entire design process directly influences the quality of technical artifacts and the overall functionality of algorithmic models. As a result, this impacts their performance in various contexts and affects their adoption. Numerous examples of poor algorithmic design illustrate this reality, often referred to in terms of ‘algorithmic bias.’ This occurs when algorithms reproduce and amplify societal biases present in their datasets, often disproportionately impacting minorities and vulnerable segments of the population (Buolamwini and Gebru, 2018; Noble, 2018; Prates et al., 2020).

Despite the challenges of translating ethical principles into practice (Morley et al., 2023; van Hilten et al., 2025), EbD represents a commitment to mapping and assessing the relevance of ethical issues and incorporating mitigation measures across all stages of technology design. It allows us to proactively recognise and mitigate ethical issues and risks in advance and from the start of technology design processes, from product ideation to data collection, model training, and deployment. When combined with context-specific frameworks such as ELSA (which is introduced in 1.3), EbD allows us to question the extent to which the benefits of a particular technology may

endanger fundamental human values at the heart of responsible innovation practices negatively impact particular communities or lack compliance with existing regulatory mechanisms (Juhász et al., 2025; van Hilten et al., 2025). At the same time, incorporating EbD practices into the multi-stage technology design process presents an opportunity to distribute ethical responsibility among the various teams and stakeholders involved. This approach fosters discussions that could expand the concept of ethics beyond the typical understanding of ethics-as merely regulatory compliance (Morley et al., 2023).

In AI-CODE, implementing EbD is crucial for developing tools and services that help media professionals to combat disinformation because journalistic work depends fundamentally on public trust. Thus, any AI-assisted intervention for identifying or countering disinformation must be designed to avoid unintended biases, overreach or impacts on editorial independence. Disinformation itself necessitates careful ethical considerations such as what constitutes disinformation, how can it be encoded in automated detection systems, and what level of information should be included in user-centric explanations. Furthermore, AI-CODE harnesses the expertise of interdisciplinary experts working on the project along with privileged access to stakeholders and future users of these tools to implement ethics principles in a practical way. This report will clarify in subsequent sections that AI-CODE draws on the AI-HLEG's *Ethics Guidelines for Trustworthy AI* (European Commission, 2019) as a starting point, and has, through interdisciplinary dialogue with project partners and stakeholders, identified shortcomings (particularly with regards to applying ALTAI to generative AI) that require the update and extension of the ALTA framework.

1.3 Addressing shortcomings of EbD in AI-CODE

Overall, EbD is an important element of the approach to ethics in AI-CODE. However, some limitations of EbD seem persistent. For example, the ideal scenario would be to substantiate EbD frameworks with a contextual assessment and mapping of relevant ethical design requirements (Brey and Dainow, 2023). In reality, organisations struggle to mobilise the resources and expertise required to conduct such assessments (Morley et al., 2023; van Hilten et al., 2025), and AI-CODE might not live up to this ideal either.

While AI ethics principles provide a broad framework, the ethical requirements for AI systems are quite specific and depend on the context in which they are deployed, as well as the particular needs of different stakeholders. For example, racially segregated data is often considered unethical, but it can be essential for providing appropriate healthcare. The relevant domain for AI-CODE is journalism, which is characterised by strong professional ethics. Journalistic ethics must be considered during the service development process to prevent ethical problems in the deployment phase.

Nevertheless, authors have noted that translational tools that support EbD, such as ALTAI, often lack domain-specific knowledge and ethical elements. Moreover, scholars have pointed out that EbD might ultimately limit ethical considerations to the design stage of AI systems and, in this manner, unintentionally overlook the ethical dimension of the entire AI lifecycle, particularly the post-design stages of AI deployment, monitoring and evaluation, included in ALTAI (van Hilten et al., 2025).

For this reason, some have argued for operationalising ethics principles through a context-specific assessment of the ethical, legal and social aspects (ELSA) of AI (van Hilten et al., 2025), which we

adopt in AI-CODE, as well¹. Initially developed to address the implications arising from genomics research (Zwart and Nelis, 2009), ELSA has become a popular framework for assessing emergent tech-related issues which are not primarily – or at least, exclusively – technical in nature. Because ELSA tends to be conducted by non-technical specialists (e.g. lawyers), pursuing ELSA in isolation can exacerbate the abstraction of ethical and legal principles from the actual technical design of these systems (Juhasz et al., 2025). By leveraging the proactive, design-centric and systematic approach of EbD in combination with the context-specific mapping enabled by ELSA, we can foster a more responsible approach to developing trustworthy AI systems and tools. We thus utilise this combination in AI-CODE.

In the project, we have begun with mapping ethical issues of concern (section 2.1), and providing (abstract) ethical design requirements based on the work of the AI-HLEG (section 2.3). The ALTAI tool supported this operationalisation. To strengthen the AI-CODE EbD practice, the consortium further relies on the expertise of an Ethics Advisory Board (EAB). The EAB, which is periodically briefed on the product development process of our different tools and services, providing input and probing into ethical design practices, in the vein of the “ethics-as-a-service” model (Morley et al., 2021).

1.4 Deliverable outline

This deliverable provides a detailed overview of the activities related to T9.3 on ethics during the first two years of the project. T9.3 encompasses various activities and includes two deliverables: D9.5 Ethics Design Requirements (due in M24) and D9.6 Ethics Handbook (due in M36). Additionally, the three annual EAB meetings are also included in T9.3, as shown in Figure 1. This deliverable focuses on the EbD activities of the initial years which are shown at the bottom of the diagram and introduced in the previous sections.

In light of this introduction to EbD and the rationale for adopting this approach in AI-CODE, the

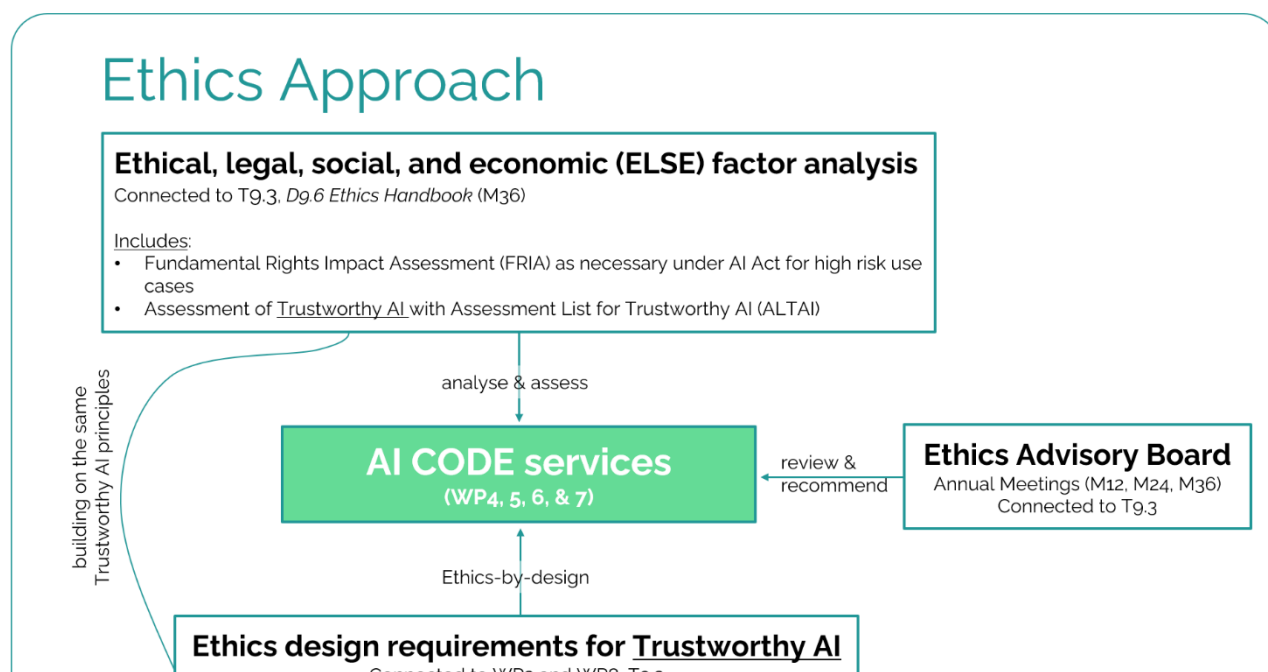


Figure 1. Diagram of Ethics Approach in AI-CODE

¹ This will be discussed at length in D9.6 Ethics Handbook (M36).

following sections explore how EbD was operationalised in the project. It starts by describing how we have mapped ethical areas of concern (section 2.1) and discusses the contribution of AI-CODE's Ethics Advisory Board (EAB) in this process (section 2.2). The ethical design requirements underlying AI-CODE are specified in section 2.3, with particular emphasis on the Ethics Guidelines for Trustworthy AI and the Assessment List for Trustworthy AI (ALTAI). The report also discusses the results of the first round of ALTAI assessments conducted for AI-CODE services (section 3). The final section of this deliverable reflects on the EbD method implemented in AI-CODE as well as the challenges and limitations faced during the ALTAI assessments (section 4.2). The document concludes with a look ahead at future steps, particularly in view of the ELSA analysis that will be conducted next year.

2. Ethics-by-design in the project

The activities of the first year included the following steps: 1) mapping 'areas of concern' in the first year with the consortium and the Ethics Advisory Board (EAB), 2) the implementation of advice by the EAB in the beginning of the second year, and 3) the proposal and selection of ethical design requirements in conjunction with WP3 based on the ethical framework of Trustworthy AI, and 4) the first ethical assessment of the services. This section describes these steps in more detail.

2.1 AI-CODE Services

Before diving into the ethics-by-design activities, the AI-CODE services shall be described briefly. The first three services are clustered together as data-driven services, and the other three services are clustered as user-driven services. The data-driven services focus on assessing misinformation content, whereas the user-driven services focus on teaching, empowering, and supporting journalists in their interaction and use of generative AI.

2.1.1 Data-driven services

Service 1 – Disinformation Detection in Next-Generation Social Media (developed in T4.1/T5.1) – includes two key modules. 1) **The Fediverse Network Moderation Tool** module identifies potentially harmful servers in the Fediverse network by categorizing banned servers based on keywords and providing an updated list through an intuitive interface. 2) **The Mastodon Credibility Service** module analyses the credibility of information on Mastodon instances using data from message reshares, instance bans, and content analysis. Overall, the goal of this service is to promptly provide journalists and fact-checkers with a sense of the reliability of the Fediverse server. By analysing message resharing patterns, moderation actions, and flagged content, AI uncovers coordinated behaviours and enables the automated detection of misinformation clusters, the tracing of information flows, and the identification of key players in dissemination networks.

Service 2 – Trustability and credibility assessment for content and sources credibility and artificiality scores from content found on Fediverse_ (developed in T4.2/T5.2) – utilises AI-based methods to return credibility and artificiality scores based on explainable assessments for content found on Fediverse microblogging platforms. This service primarily integrates a bot detection tool, network graph on how a post spreads, and a hate speech detection tool (from service 6).

While the second module of service 1 analyses servers in platforms like Mastodon and BlueSky, service 2 operates additionally on the level of posts.

Service 3 – Media Annotation and Asset Management service (developed in T4.3/T5.3) – allows media professionals to upload images and videos (media assets) to determine whether they are authentic or synthetic. MAAM provides annotation (labelling) and analysis of these media assets. It integrates forgery detection, explainability features, and user profiling for disinformation analysis.

2.1.2 User-driven services

Service 4 - Interactive Simulation and Coaching & Personal companion (developed in T6.1/T7.1) – includes two modules. 1) **The Interactive Coaching and Simulation Module** (ICSM) helps media professionals understand the principles of generative AI (e.g., Large Language Models) and their associated risks by teaching users how to construct prompts and interact with systems to produce trustworthy content. It includes guidance on reliability analysis of LLM outputs and verification of their accuracy. 2) **The Counter-Narrative Recommendation Module** (CNRM) offers on-demand suggestions for claim verification. Users can input an online URL and a claim based on this source. The module then analyses whether the claim is accurate, and highlights sentences from the source to substantiate the verdict. It helps users develop counter-disinformation strategies and supports claim verification workflows.

Service 5 – Transparency Service for AI Model Cards (developed in T6.2/T7.2) –provides a standardised structure for creating and managing AI model cards, including performance metrics, model limitations, and ethical considerations. It offers a user interface and software development kit for card creation and management. Additionally, the service provides varying levels of detail within model cards to cater to both technical experts and non-technical stakeholders.

Service 6 – Personal Companion for Understanding Disinformation (developed in T6.3/T7.3) - assists journalists and fact-checkers in combating the spread of dis- and misinformation by offering a range of features, including assistance in writing credible articles and checking content validity by highlighting logical fallacies, hate speech, and debunked claims in texts.

2.2 Mapping areas of concern

As a first step of the ethics-by-design approach, areas of ethical concern were mapped for each AI-CODE service and some overarching issues were collected. The purpose of this was threefold. First of all, it was an educational and awareness raising exercise. Ethics-by-design requires that designers of technology are aware of ethical pitfalls and understand the seriousness of these. Therefore, the purpose of mapping was to engage the entire consortium in a critical reflection on their services and potential ethical issues brought up in the literature around AI tools to fight misinformation. Second, the mapping serves as a basis for prioritisation. There are various time and resource constraints at play in developing an AI system which can overwhelm and practically undermine ambitions for implementing ethics (Carillon, 2024). On the other hand, ‘ethical AI’ can become overly demanding. As discussed above, the ideal implementation of EbD is very demanding and requires context-sensitivity. In order to ensure that the ethics approach in the AI-CODE project remains within implementable bounds, we have deemed important to prioritise the most pressing ethical issues. Therefore, the mapping has placed a special emphasis on the specific issues related to the AI-CODE services rather than addressing the broader context of the project. The goal has been to make the

insights emerging from the ethics literature review and mapping exercise usable in the design and development of the services. Third and finally, another purpose of the mapping was to facilitate the discussion at the first annual meeting of the Ethics Advisory Board.

The mapping began with identifying issues relevant for AI-CODE based on the exploratory literature review (section 2.1.1). These were visually mapped in a Miro Board (Annex I), consulted with the independent ethics advisor of the project, and further expanded with additional input from the entire consortium (section 2.1.2).

2.2.1 Exploratory literature review to identify general ethical issues

The goal of the exploratory literature review, carried out in the first months of AI-CODE (Q1/2024), was to identify overarching ethical concerns related to AI and misinformation. Initial research papers were identified by searching for publications on Google Scholar with the key words ‘AI against misinformation ethics’. From there, further relevant papers were identified through references. The scope of this literature review was limited, since the goal was an exploration as a basis for defining concrete issues related to the AI-CODE services. What the exploration found was that most papers discuss the risks of AI to generate and spread misinformation. However, few articles consider the ethical issues related to leveraging AI to detect and address misinformation – the concrete ambition of the AI-CODE project. Those which apply this focus on misinformation detection are highlighted here.

For example, Bontridder and Poulet (2021) tackle the overall ethical issue of defining mis- and disinformation for the purposes of algorithmic content moderation in a way that their operationalization does not infringe on the freedom of expression. With regards to *ex ante* filters and the blocking of content, they argue that these filters require high levels of explainability and transparency so that they can be monitored for over-restricting the freedom of expression. Limited accuracy of detection models and upload filters can lead to discrimination and reduce the diversity of represented opinions, thus negatively affecting democracy (Bontridder & Poulet, 2021, p. 11). The authors further argue that if the algorithm prioritizes AI-generated content, it can result in problems for media pluralism which relies on the work of different media outlets (produced by human writers and editors) reaching audiences (ibid., p. 11). Therefore, the accuracy of AI-CODE services for identifying AI-generated media items was highlighted as a priority.

Zhou et al. (2023) also discuss the role of accuracy in fake news detection. They find that the accuracy of readily available disinformation detection tools decreases for AI-generated fake news. They argue that this can have ethical implications regarding the kind of news that are considered by the public as true and that security risks can emerge from AI-generated misinformation. While they advise to work on improving detection tool, they warn against an ‘arms race’ between researchers and developers of detection tools and bad actors. Further, they discuss the ethical concept of ‘responsibility’ and assign a clear responsibility to journalists to inform the public about AI capacities and on the persuasiveness of AI-generated misinformation (Zhou et al., 2023, p. 16). Therefore, of relevance for AI-CODE services, is the importance of empowering journalists to take on this responsibility.

Responsibility is also discussed by Gillespie (2023) but at another level of analysis. He raises the question of whether independent and academic researchers should be invested in ‘fixing’ the problem of misinformation, or whether this should be an issue for social media platforms to tackle, as their recommender systems and business model are the underlying causes. He argues for the

latter. While additional regulation of social media might be another path of action, the current problem of the spread of mis- and disinformation on social media should not be neglected, as real harms arise. Providing actionable services for today's problems is the goal of the AI-CODE project.

Allein et al. (2023) discuss the issue of profiling by algorithms designed to detect fake news spreaders. Their ethical anchorage lies in the fact that “people have the right not to be subject to decisions based on automated, ‘human-out-of-the-loop’ processing that assesses their personal aspects — especially when those decisions produce legal effects concerning them” (ibid, p. 1). The authors propose an algorithm which avoids profiling, thus demonstrating that this can be technically achieved. The issue of profiling is considered in the mapping as a practice that needs to be avoided.

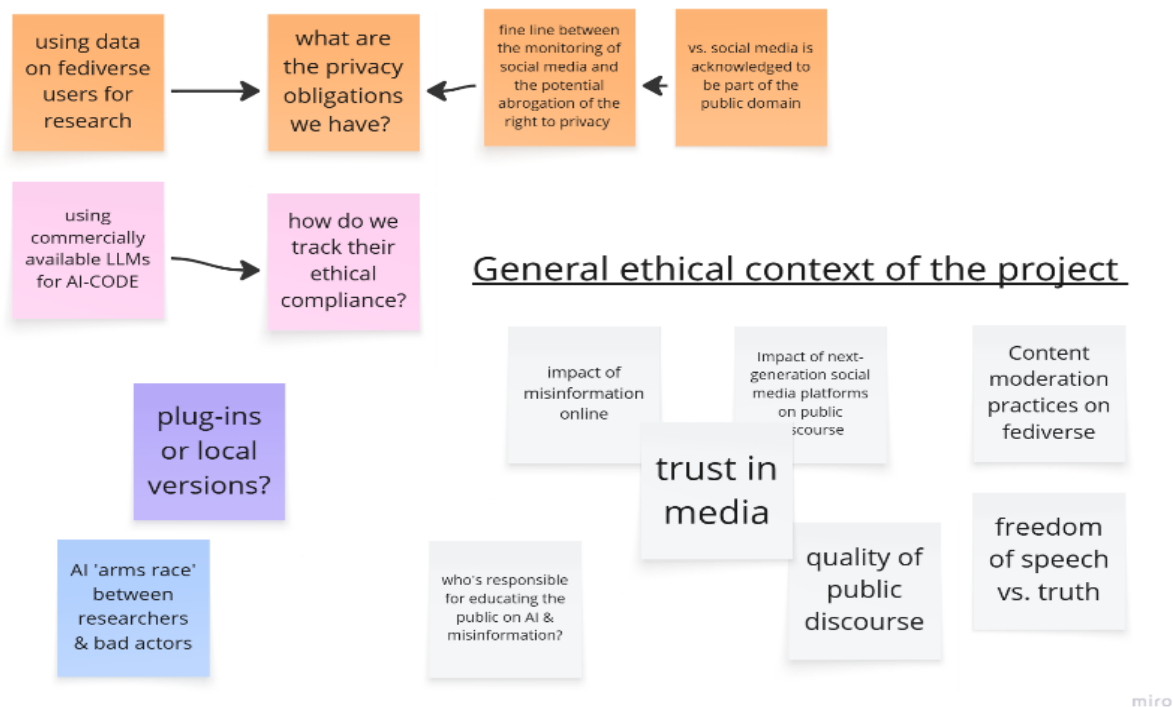
Cartwright et al. (2023) consider the use of different machine-learning algorithms for classifying data as information or misinformation. In terms of ethical aspects, they address the existing controversy about content moderation: either large platforms are not doing enough to protect users or they are going too far in censoring free speech. Another aspect raised by the authors is the use of social media datasets for research, which seems to infringe on privacy rights (ibid.). Against this, they hold the widely acknowledged fact that users are aware of posting in the public domain. This ethical issue has even more importance in the context of Fediverse data (Lisker & Mihaljević, 2025) and will be tackled in the second annual EAB meeting (M24) and will be reported in D9.5.

2.2.2 Visual Mapping of ethical issues

Based on the exploratory literature review and analysis of the services in light of the principles of the *Ethics Guidelines for Trustworthy AI*, ethical issues were collected in an unstructured way at first, before they were added to the shared online whiteboard for a visual representation. In one part of the board (see Fig. 1), the overarching or general issues are presented, while the other part maps these issues to the AI-CODE services where they are of most immediate relevance. The content in Fig. 1 establishes the ethical context of the project and open questions that are relevant for all services. The emphasis for the initial mapping was related to AI, which led to the exclusion of service 1, since it does not utilise machine learning (ML) techniques. Ethical concerns related to service 1 were collected through other channels, though, and will be discussed in this Deliverable (see section 2.3.2.1, section 3.1, annex II and annex IV).

After a first gathering of ethical issues, CEPS shared the online whiteboard with the independent ethics advisor of the project, who made additions based on the *Ethics Guidelines for Trustworthy AI* by AI-HLEG, before the consortium provided their feedback as well. The online whiteboard of each service can be found in Annex 1 or online via this [link](#).

Non-service-specific areas for ethical consideration



miro

Figure 2. Mapping general ethical areas of concern

The visualisation has been used in the background to prepare the first EAB meeting which focused on providing feedback on each service. It is furthermore a resource that has focused the attention of CEPS in implementing their ethics-by-design approach. With an eye towards the ELSA analysis which will be conducted to counterbalance the shortcomings of EbD, the mapping serves as a first step. Rather than doing ELSA in isolation from EbD, the project attempts to implement these approaches in lockstep to allow for the cross-fertilisation of methods, discoveries, and perspectives.

2.3 The Ethics Advisory Board

As mentioned in section 1, an ethics advisory board (EAB) can support the distribution of responsibility for implementing EbD and can contribute to an ethics-as-a-service approach. AI-CODE's EAB meets annually and falls under T9.3, led by CEPS. The role of the EAB is to oversee and improve the ethical practices in the project by identifying gaps in the current approach, as well as sharing best practices and examples on how to address these issues. The second EAB annual meeting (M24) will have taken place by the time that this deliverable is submitted (M24) but cannot be reported in this deliverable due to the close timing of both. The second meeting focused on two ethical issues that have come up bottom-up and across services. The first issue is the ethics of using Fediverse user data for research purposes, which many Fediverse users oppose (Lisker & Mihaljević, 2025). The second issue relates to the ethical issues that LLMs carry into downstream applications, where many aspects can no longer be addressed, such as those relating to the training practices (discussed in section 4, as well). This will be described in D9.6 Ethics Handbook. The discussion below is thereby restricted to the first EAB meeting in M12.

2.3.1 The first annual EAB meeting (M12)

The objective of the first annual meeting of the EAB (see D10.1 for the composition of the EAB), which took place in November at the CEPS office in Brussels, was the identification of the most pressing ethical issues related to the early conception of the services that should be monitored and addressed as the project continues. The EAB meeting began with a round of introductions, followed by a short presentation of the AI-CODE project overall and clarification of the role of the EAB. The AI-CODE EAB is not an ethics review board that deals with ethical clearance for research activities. Instead, it works on a more conceptual level, engaging with the ethical issues related to the services as future products. Each service was presented and follow-up questions were clarified before the EAB members started raising ethical issues related to each service.

To ensure that the feedback from the EAB was translated into concrete design practices and decisions of the AI-CODE services, CEPS has conducted one-on-one follow-up meetings with each service team to communicate the feedback and agree on concrete measures to be adopted to mitigate the ethical risks pointed out in the discussion. This exercise allowed for the difficult process of operationalisation (Morley et al., 2023) through interdisciplinary exchange. CEPS contributed with social science and ethical expertise, while service developers have technical and design knowledge, which is necessary for implementing the feedback on ethics. Measures were proposed and reviewed in a collaborative manner. The resulting table with agreed upon mitigation measures for each service is found in Annex II of this deliverable.

A positive learning experience was that many ethical concerns can be addressed with relatively simple technical design elements. For example, concerns about over-reliance and human oversight can partially be addressed by providing optional buttons to ‘learn more’ about the underlying AI systems. The impact of the measures implemented in response to the EAB feedback has become apparent in the positive assessment of implementation of ethical design requirements in the mid-project assessment (see section 3 of this report). Section 2.2.2 recaps the service-specific feedback raised by the EAB, which informed the further design and development of the services.

2.3.2 Service-specific feedback from EAB (M12)

The first EAB meeting was a lively discussion with complementary expertise in philosophy, epistemic agency, digital policy, gender and security. Many clarification questions about the services were asked. When the consortium representatives were unable to answer them, they were discussed with the service leaders in follow up meetings. Overall, the EAB found more conceptual challenges with the data-driven services developed in WP4&5 than with the user-driven services (WP6&7) that emphasise educational awareness-raising. Table 1 include short summary of the service-specific feedback, which is spelled out in more detail below.

Table 1. Summary of EAB feedback

SERVICE	MAIN FEEDBACK
1 Disinformation detection in the next-generation social media (T4.1/5.1)	- Review the reliability of instance classification based on the moderation decisions - Reconsider including account names in analyses to protect users’ privacy
2 Trustability and credibility assessment (T4.2/5.2)	- Consider providing an artificiality and a credibility score rather than aggregating in one

3 Media Assessment and Annotation Management (MAAM) (T4.3/5.3)	- Indicate clearly reliability of the service to users to prevent over-reliance - Clarify terms of service, especially how user uploads are (or are not) stored - Make transparent how the models perform across different cases
4 Interactive Coaching and Simulation Module (T6.1/7.1)	- Clarify how we understand the goal of 'responsible' use of AI of this service
4 Counter-narrative recommendation tool T6.1/7.1)	- Clarify the ways in which the claim identifier operates (e.g. how is 'claim' defined) and what its limitations are
5 Transparency Services for AI-CODE Services (T6.2/7.2)	- Monitor if the use of a chatbot to fill in the model cards introduces biases - Write ethical 'risks', rather than 'considerations'
6 Personal Companion for understanding disinformation for media professionals (T6.3)	- Adjust wording to create appropriate user expectations - Consider expanding the database for claim matching beyond individual organisation

2.3.3 Service 1: Disinformation detection in the next-generation social media (T4.1/5.1)

The first service raised a few clarification questions, including whether its aim is to ban a Fediverse instance or a post. It was clarified that this service aims to provide information, rather than exercising any specific content moderation actions. It was specified that it operates on the level of instances, not on the level of individual posts or content. It was received well that the goal of this service is not to identify true or false information but to pick up on credibility signals that the community provides. However, it was pointed out that this might cause issues because it can function as a 'halo effect' and cause over-reliance. For example, an instance on the Fediverse might seem overall credible based on the analysis of this service. This might influence the perception of the service user that all content coming from it is reliable.

Another important question discussed at length was whether the service developer would review information on why instances are banned and which keywords were assigned to them (e.g. hate speech, racism, etc.). The EAB pointed out that the classification of instances in content moderation decisions is not beyond doubt. For example, one can imagine revenge banning of instances. The recommendation for the short term is to randomly sample instances to confirm whether they are correctly classified. In the long term, the service team should clarify to the service user who takes responsibility for how instances were classified and who can be held accountable for a misclassification of instances.

The EAB members also discussed the information that one can receive through instance characterisation, such as 'most prominent users' and 'most prominent posts'. For the EAB, this raises the question of how 'prominence' is defined. They recommend identifying and showcasing the most prominent post and user of the last three months (or a similar timeframe), rather than an 'absolute' understanding that takes into account the entire existence of the instance, given that misinformation comes in waves. They also raised the question of the added value of knowing the

most prominent user. They are likely to use a pseudonym anyway, and to single out a user raises privacy concerns. Particularly in publications, the account names of the most popular users should not be included to prevent negative consequences.

The EAB also discussed the question posed by the consortium on using publicly available user information in the context of AI-CODE research. The EAB agreed that it would be worthwhile to address this question because the data is not anonymised (i.e., the service identifies specific users). It appears permissible to process this data. However, we should not publish the names of the most important users because that is personal information.

2.3.4 Service 2: Trustability and credibility assessment (T4.2/5.2)

The service developed under T4.2 and continued in T5.2 consists of multiple functionalities and credibility indicators. The EAB pointed out that the service investigates two separate issues. i) Is the content or post ‘artificial’, i.e. AI-generated? ii) How credible is the post origin and the accounts that engage with it? The EAB suggested that transparency is enhanced if we provide an ‘artificiality score’ and a ‘credibility score’ instead of using a single score.

For the bot-identifier, the EAB was curious about how many false positives and negatives there are. Overall, the question is whether we have enough confidence in the results of this tool to rely on it. The EAB recommends informing the user about the appropriate level of reliance.

As the consortium had asked at some point whether this service could be considered a social scoring algorithm prohibited under the European AI Act, this question was brought to the EAB. The EAB leaned towards saying that it could only be considered a user account scoring, a conclusion with which they also did not fully agree. It does not target overall social behaviour and is not government administered. The fact that it focusses on assessing credibility of posts and less so on the identity and punishment of the poster further distinguishes it from social scoring.

2.3.5 Service 3: Media Assessment and Annotation Management (MAAM) (T4.3/5.3)

For the MAAM service, an important question raised by the EAB was what happens to the (potentially sensitive) images that the user uploads. The EAB asked for additional information on where this user data is stored, whether it is used for further training of the system, and whether the user can delete the images permanently. The last question brings into play the ‘right to be forgotten’ of the GDPR and becomes particularly pressing in the context of journalism, where media professionals handle sensitive data from their sources. The EAB recommends that the user should have an option to delete any content they upload for annotation and analysis.

Further, they warn that we should not assume that MAAM will be used exclusively in a professional context. Media professionals are likely to also use it for private purposes. The right to be forgotten, i.e. the ability to permanently delete the data one uploads, should be implemented. The EAB recommends that when a user uploads content for analysis, they are shown the terms of service, which should include the following details:

- Information to enhance explicability of the analysis results and explainability of the process, so that the media professionals have a minimum understanding of why the system decides one way or another.
- Information on the weaknesses of the system. For example, typically faces are more reliably detected than e.g. synthetic images of buildings.

- Information on whether user data is used for further training or algorithmic decision making.

2.3.6 Service 4: Generative AI Interactive Coaching and Simulation & Counter Narrative Recommendation Service (T6.1/7.1)

Component 1: Interactive Coaching and Simulation Module

The Simulation tool was perceived by the EAB as being particularly useful, given the need for more (public) education on generative AI. Particularly, the EAB considered that we need to train people to ask where sources come from and to understand the blind spots, biases, and hallucinations of generative AI. For this reason, this educational tool was deemed very useful. The main question raised for this service aimed at how we define responsibility, or ‘responsible use’ in the objective of the service, which is stated as follows: ‘The interactive coaching and simulation module (4.1) aims to help journalists use GenAI tools responsibly while adhering to professional values and standards.’ The EAB recommends that we adopt a multidimensional definition of responsibility in relation to AI which reflects the different parties that engage in the development of AI.

Component 2: Counter-narrative Recommendation Tool

For the sub-service of the counter narrative recommendation tool, the first EAB clarification question was how a counter narrative is recommended and which data is used. It was clarified that this tool uses the data provided by the service user and builds on an LLM to generate a useful counter-narrative recommendation.

The EAB had several questions revolving around the claim-identification process. They asked for clarification on how ‘claim’ is defined in the service. The AI-CODE members present at the meeting suggested to understand it as the basic unit of information. From here, they wondered further whether the service distinguishes between false claims and true claims, i.e. factual statements, already in the identification process. The EAB understands the challenge of making such a distinction, but the EAB emphasises that there is a difference between fact and claim, especially since digital environments and post-truth movements seem to turn facts into claims.

The EAB also found the question of how the identifier deals with weaker claims to be of ethical relevance. For example, sometimes claims are expressed as questions, using modal verbs, or in terms of ‘might... possibly’. What kinds of conditionals can the claim identifier handle? This question revolved around the complexity of identifying a claim. Doubts were raised about whether we can promise that all claims are being identified since there are many utterances that cannot easily be distinguished into claim vs truth. Related to the questions above, claims can be formulated with differing levels of confidence. From an epistemic point of view, it might not be a strong claim, but it can still impact users who engage with it.

The EAB recommends clarifying the exact ways in which the claim identifier operates and what its limitations are. The user requires this information to interpret the results and adjust their reliance on the service to an appropriate level.

2.3.7 Service 5: Transparency Services for AI-CODE Services (T6.2/7.2)

It was explained that the transparency service should provide different levels of technical detail in model cards that are generated for different audiences. The idea to provide both a detailed text and something like an executive summary was well received. The EAB suggested that there would be no significant trade-off between transparency and usability if different levels of detail were made available. In deciding the level of detail, the EAB recommended making it dependent on how much the user wants to know.

Summary levels were deemed useful, and it seems to be permissible to use a chatbot to this end. However, if advanced chatbot functions are integrated in the future, this chatbot will itself require a model card to be transparent. They also highlighted that the use of a chatbot to fill in the model card invites biases, and thus we risk creating unreliable model cards. This should be monitored in the further development of the service.

Subsequent discussions focussed particularly on the ‘ethical considerations’ that are documented in the model cards. The EAB recommended re-naming it to ‘ethical risks’, because otherwise they might invite very long texts on ethical benefits, when we are mostly interested in the ethical risks that should be considered when deciding on whether to use this model. The EAB recommended presenting the risks at a glance, without losing the detailed explanation. They proposed to integrate a drop-down list here as well. For example: The cell specifies five ethical risks to consider. Each one can be opened in more detail. In terms of ethical risk documentation, the EAB recommended including the source of bias in training and evaluation datasets explicitly in the cell of ethical risks, and not just when the dataset is reported.

2.3.8 Service 6: Personal Companion for understanding disinformation for media professionals (T6.3)

The first point raised with regards to this service relates to the description stating it aims to: “assist young journalists to write articles without any logical fallacies”. The EAB was concerned that this might be over-promising. While logical fallacies can be categorised into different types, it is very subtle when they occur in natural language. It also raises the same issues around breaking down a text into claims (which has been discussed for service 4 in detail) and to identify premises and conclusions based on that. The EAB requested additional information on how this works exactly and how reliable the tool is. It further recommended reviewing the language to avoid over-promising.

Another issue was raised regarding the wording in the above quoted sentence: It seems unnecessary to specify that this is for ‘young’ journalists. Age does not protect one from logical fallacy.

The organisation-specific database was noted to be useful from a privacy perspective and responsibility clearly remains with the organisation. However, in the long term it seems beneficial to have the option to share data with other trusted organisations, possibly protected with a system of cryptographic hashes. The EAB noted that, while this is not yet relevant for this phase of the project, it can be once we aim to place it on the market. About the debunked database, the EAB pointed out that its quality depends on the critical literacy of media professionals. They approved of the common fact-checking practice of recording the reasons for ‘debunking’. This functionality was deemed useful to enhance critical thinking and trust in the system.

2.4 Ethical design requirements

Beyond mapping issues and consulting an EAB, adopting an ethics-by-design (EbD) approach also requires the concrete consideration of ethical principles. As highlighted in section 1.2 of the introduction, EbD relies on commonly agreed ethical principles, assesses how the designed product aligns with these, and consults different stakeholders. In AI-CODE, our EbD approach is informed by the work of the EU High-Level Expert Group on AI (AI-HLEG). Most notably, we draw on the *Ethics Guidelines for Trustworthy AI* put forward by the AI-HLEG, as well as the accompanying *Assessment List for Trustworthy AI* (ALTAI). Where the Ethics Guidelines provide an overarching framework for a principle-based development of ethical AI systems, the ALTAI allows for evaluating the trustworthiness of this process and the resulting AI systems. ALTAI is discussed in greater detail in section 2.5.

With regards to the selection of the AI-HLEG guidelines as the core framework for AI-CODE, this has been motivated by a number of different reasons, namely:

- i) **The scope of the Ethics Guidelines is broad.**
Rather than focussing on one specific application or economic sector, the principles are applicable to a wide range of possible use cases, which allows us to apply them in the AI-CODE context.
- ii) **They are practice-oriented.**
Compared to an academic framework, these guidelines have been part of a policy initiative, aimed at wide uptake of the principles in the development of AI in the EU.
- iii) **They are impactful.**
The Ethics Guidelines have been influential for AI governance and ethics. The review of ethical guidelines by Corrêa et al. (2023) shows that there is overall convergence around ethical principles, such as those suggested by the AI-HLEG.
- iv) **They are based on expert consensus.**
While the reception of the guidelines has not been without critics (Ryan, 2020), the fact that 50 independent experts have worked together to write these guidelines, reaching broad consensus, is notable, as well as the aforementioned international convergence.

The AI-HLEG Ethics Guidelines comprise seven core principles for trustworthy AI which are introduced in more detail in the following section 2.4.1, alongside the context of its development. These principles have subsequently been translated into specific design requirements that address the context-specific needs of journalistic practices at the heart of the AI-CODE project, facilitating their embedding in the AI-CODE services, as described in section 2.4.2.

2.4.1 The AI-HLEG Ethics Guidelines for Trustworthy AI

The *Ethics Guidelines for Trustworthy AI* were published in 2019 on behalf of the AI-HLEG by the European Commission (EC) (European Commission 2019). The European Commission set up the AI-HLEG in 2018 to lay the foundations for a European approach to AI. It was one of the first AI governance initiatives and set the tone of the European strategy. In light of different scandals breaking related to the use and misuse of AI systems, the European approach foregrounds ethical AI. While this was not a regulatory initiative, the AI-HLEG's Ethics Guidelines have played a role in shaping the European AI Act.

The AI-HLEG states that an AI system can be deemed ‘trustworthy’ if it is lawful, ethical, and robust (European Commission, 2019, p. 2), as also explained in D3.1. This is further spelled out in seven key requirements that need to be met so that the development, deployment and use of AI systems can be considered trustworthy, namely (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.

The role of these guidelines in AI-CODE has been to define what we mean by ‘ethics’ in ‘ethics-by-design’ with regards to the different AI systems. The principles articulate the expectations for the service developers and external stakeholders, including the identified user groups. The framework allows for identifying the standards for ethical AI that can be utilised for concrete analysis of the stakes at play in the development and deployment of the AI-CODE services. One shortcoming is that the principles are abstract and require specification. We have addressed this issue by drawing on the seven ethical principles and using the accompanying assessment list for trustworthy AI (ALTAI) (see section 2.5) to define concrete design principles for AI-CODE services.

2.4.2 From Guidelines to Design Principles

Notable with a principles-based approach is the challenge of operationalisation mentioned in section 1.1 already (Morley et al., 2023). Principles require context-dependent interpretation and are better at guiding at an abstract level, while remaining ambiguous in concrete situations. For example, transparency in the journalistic context is important especially when it comes to making transparent how a news outlet or a story is funded. However, transparency with regards to the sources of journalistic reports should be balanced with protecting the privacy of their sources. Thus, transparency is required and an imperative in some cases, but not an absolute requirement for other instances within journalism.

A first step then for our EbD approach was to translate the abstract trustworthy AI principles into high-level design requirements for AI in the journalistic context. The ethical design requirements for AI-CODE were drafted based on these seven trustworthy AI requirements, by asking the question: “What is needed in the design of the system to be able to say that one has fulfilled this requirement?” Since the ethical design requirements have already been presented in D3.1, in this deliverable they are included in Annex III. Each service has selected a number of principles that were considered a priority and worked towards their implementation, which was documented in the WP4/WP6 deliverables. In the project, we have found the use of an ethical assessment tool to be a helpful guide to understand better what needs to be done concretely to implement these ethical design requirements. ALTAI is the appropriate method to this end in AI-CODE because it is based on the same *Ethics Guidelines for Trustworthy AI* as the design requirements specified in WP3.

2.5 Evaluation with Assessment List for Trustworthy AI (ALTAI)

The *Assessment List for Trustworthy AI* (ALTAI) was developed by the High-Level Expert Group on AI (AI-HLEG) on the basis of the *Ethics Guidelines for Trustworthy AI*. It is clearly stated that it “is intended for self-evaluation purposes” and “provides an initial approach for the evaluation of Trustworthy AI” (European Commission, 2020, p. 2). The goals stated by the expert group include raising awareness of risks and supporting the implementation of proactive measures to foster ethically responsible innovation (p. 3). To this end, the AI-HLEG proposes to involve a multidisciplinary team, composed of members from within and/or outside the organisation, such as

“AI designers and AI developers of the AI system; data scientists; procurement officers or specialists; front-end staff that will use or work with the AI system; legal/compliance officers; management.” (European Commission, 2020, p. 4)

In line with the broad applicability of the Ethics Guidelines, the expert group also encourages flexibility in the use of ALTAI:

This Assessment List for Trustworthy AI (ALTAI) is intended for flexible use: organisations can draw on elements relevant to the particular AI system from this Assessment List for Trustworthy AI (ALTAI) or add elements to it as they see fit, taking into consideration the sector they operate in. (AI-HLEG, 2019a, p. 3).

It is in this spirit of flexible use that ALTAI was adapted to fit the concrete needs of the AI-CODE project, while maintaining the goals of awareness raising and guiding the implementation of proactive measures.

2.5.1 Method in the project

While the AI-HLEG identifies stakeholders that should be consulted for an ALTAI interview, it does not specify the timing of the assessment in the AI lifecycle. The ideal timing for the questions varies. For example, data protection considerations should be taken into account from the conceptualization on, whereas details on cybersecurity – and thereby safety – require further progression towards the infrastructure of the final product. Due to this uncertainty, we conduct ALTAI-based interviews in M18 and M36, while the content for these differs slightly. The staged assessment also allows tracking progress over time and allows for the iterative implementation of ethical principles. Furthermore, identifying gaps in the current approach at the halfway point of the project allows for proactivity in implementing additional measures to enhance the ethical compliance of the services. We are not only placing safeguards a posteriori on a finished project, but shaping the services throughout their active development.

Since we are planning to re-assess the services at the end of the project, we have reviewed all 150 ALTAI questions in advance, to prioritise those that fulfil the following two requirements:

- 1) they are appropriate to ask in M18 of the project, and
- 2) they can be answered from the perspective of the developing team, rather than from future users.

Based on the first requirement, we have decided to skip the principle *Technical robustness and safety* for the M18 assessment. The questions in this category, such as those pertaining to risk management systems and cybersecurity, gain more relevance as the services achieve greater maturity and are closer to placing on the market. At this point of the project, and since we are not handling highly sensitive personal data, we can postpone this category to M36, when higher technological readiness will be achieved.

The second requirement was applied because some questions in ALTAI relate to the epistemic faculties of users – e.g. for the oversight requirement, one question is as follows: “Is it sufficiently clear to the user that they are interacting with an AI system?” It might seem obvious to a developer that the service is AI-based, while a user might somehow overlook this fact and thus might not be sufficiently aware of it. Since users will be involved in WP8 for the user-based evaluation of services, these questions have been included in the user testing script, as explained in D8.2.

Overall, we have selected 91 out of 150 ALTAI questions for the M18 assessment. For this preliminary assessment, it was deemed sufficient to talk to the service leaders and team members from the technical side, whom the service leaders wanted to invite. With regards to the stakeholder groups proposed by the ALTAI, we thus covered designers and AI developers of the AI system, data scientists, and to some extent managers, as well. If there were other team members that should have been consulted, such as front-end and UI developers, they were consulted in follow-up emails. Through the user-based evaluations in WP8, we will also receive input from procurement officers or specialists and front-end staff that will use or work with the AI system, thus covering almost all stakeholder groups defined by the AI-HLEG, and providing feedback on the trustworthiness of the systems from the user perspective.

The interviews were conducted in M20 and M21² via Microsoft Teams and lasted between 60-100 minutes. Each interview was conducted by a research team member and a note-taker from CEPS. The interviews were recorded for note-taking purposes with the spoken consent of the AI-CODE interviewees. The recordings were deleted as soon as the notes were completed, typically within a week. The interview was facilitated with the help of an Excel sheet with the selected questions. This file was shared with the service leaders in advance and after the meeting, as well. The answers were recorded in the Excel file.

ALTAI includes two types of questions: those asking for the concrete measures that were adopted by the developers and questions that ask the interviewees to assess how they see their mitigation measures on a scale from non-existent, fully inadequate, inadequate, adequate, to fully adequate. In the analysis of the answers, we have assigned points for each adopted mitigation measure (yes = 1 point, no = 0 points, to some extent = 0.5 points) and translated the qualitative value of self-assessment into a scale of 0-4 points.

Based on the achieved points per requirement, we have created spider graphs to visualise the current state of trustworthiness based on the mitigation measures and the self-assessed level of trustworthiness. Splitting these two assessments in the visualisation served the educational goal to demonstrate when and where some services might be overestimating their trustworthiness or underestimating the progress they have made already. The filled in Excel files are available on the project SharePoint and the visualisations are featured in section 5. The results, including recommendations on improvements, were shared with the consortium at the in-person consortium meeting in M20 and once all ALTAI interviews were completed, the preliminary assessments were shared in the form of reports, uploaded to the project SharePoint. Section 3 of this deliverable presents the results, as well.

3. Preliminary results on implementing ethical design requirements

The M18 assessment with ALTAI has been a check-in on the progress towards implementing the trustworthy AI design requirements. The results are preliminary, because there is still another year left to the project and not all 150 questions of ALTAI were asked in this first assessment, such as

² The timing of the ALTAI assessment varied for each service based on the availability of the service teams. The start was delayed because several deliverables for WP4 and WP6 were due in M18. To achieve the educational goal of ALTAI, it appeared more important not to overburden the services and postpone the assessment to M20. This allowed the services to reach a higher level of technological readiness, making the assessment more impactful.

those relating to the Trustworthy AI requirement for technical robustness and safety and those that are best answered from the user perspective. The snapshots of these assessments are reported for each service in the following sub-sections. The recommendations based on the results were shared with the consortium and are included in tables in Annex IV.

3.1 Service 1 (T4.1): Preliminary ALTAI results

In so far as they are quantifiable, the results for service 4.1 *Disinformation Detection in the Next-Generation Social Media* – which provides journalists and fact-checkers with a sense of the reliability of the Fediverse server – are presented in Figure 3 below. A short discussion of the main findings will be provided, before a list of recommended measures was proposed to the service leaders (see Annex IV – ALTAI Recommendations, T4.1 Recommendations based on M18 ALTAI).

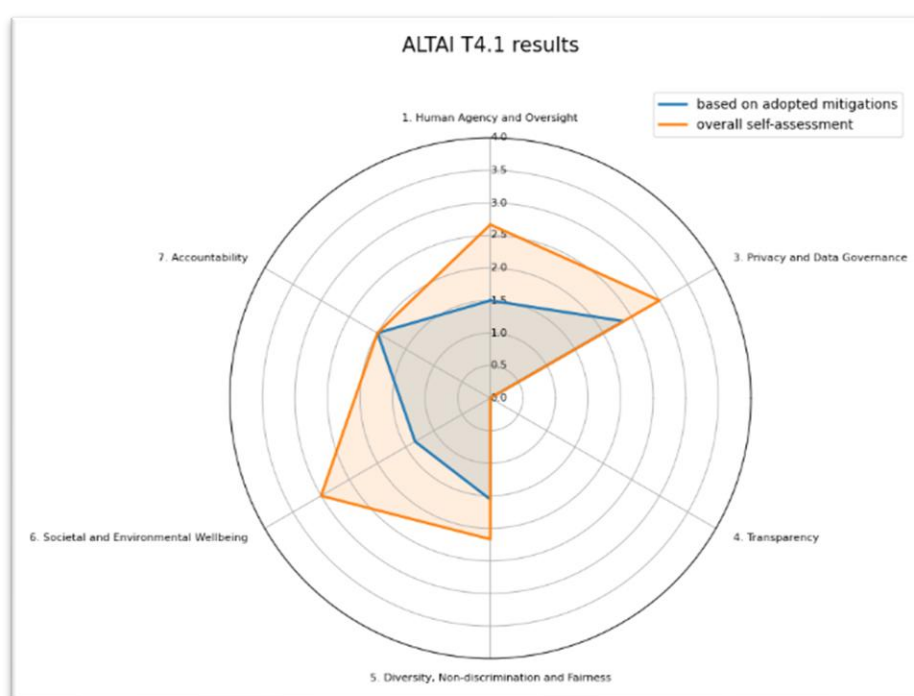


Figure 3. Spider graph T4.1 ALTAI

For 3 out of 5 requirements, the self-assessment (orange in diagram) was more optimistic than the implemented measures justify (blue in diagram). For this service, the assessment is complicated by the fact that it integrates many different models which were developed by multiple AI-CODE partners. Often, a compromise answer across service components was struck that reflects an average of the integrated service.

For this service, most improvements can be made towards **Transparency**. This can be explained by the fact that the unified user interface (UI) has not yet been developed and deployed. As is evident from the table on mitigation measures, most requirements can easily be reached when a UI is implemented which informs the user about the integrated AI models. The impact on **Societal and Environmental Wellbeing** should be addressed more comprehensively, as well. For example, the extent of the environmental impact is not fully known for this service. There is also more room to understand better how exactly this tool will impact work arrangements. However, it should be noted that there is no deskilling risk identified for this service, as it does not take over any human tasks or replace the need for critical examination of evidence.

The gap towards achieving **Human Agency and Oversight** can partially be explained by the lack of a UI, which can include a warning against over-reliance and help users overall to calibrate the appropriate level of reliance. **Diversity, Non-discrimination, and Fairness** requires, according to the *Ethics Guidelines for Trustworthy AI*, the implementation of Universal Design principles. As was stated before, the user interface has not yet been developed, thus making it impossible to reach the full score on this requirement. Further, users don't have a way yet of flagging issues to the service developers. Overall, for this requirement the understanding of the actual challenges that this service poses can be improved through additional research and testing. **Accountability** is already relatively high. The overall score was lower than it could be, as auditability is not fully ensured and there is no clear mechanism for users to communicate with the developers. Finally, **Privacy and Data Governance** should be strengthened by empowering the users of the system further, e.g. by allowing them to flag issues and withdraw consent.

Overall, trustworthiness of the service will easily be improved by the careful implementation and design of the UI. From there, additional measures (proposed in Annex IV) can be considered to further follow through on EbD.

3.2 Service 2 (T4.2): Preliminary ALTAI results

In so far as they are quantifiable, the results for service 4.2 *Trustability and Credibility Assessment* - which provides credibility and artificiality scores for content found on the Fediverse - are presented in Figure 4 below. A short discussion of the main findings will be provided, before a list of measures is proposed (see Annex IV – ALTAI Recommendations, T4.2 Recommendations based on M18 ALTAI).



Figure 4. Spider graph T4.2 ALTAI

Overall, the self-assessment (orange in diagram) produced similar trustworthiness levels compared to the level indicated by the implemented measures (blue in diagram). The only deviation occurred with a lower self-assessment of the **Transparency** requirement and a higher self-assessment of meeting the **Diversity, Non-discrimination and Fairness** requirement.

Most improvements towards the trustworthiness of the service can be made on the requirement for **Diversity, Non-discrimination, and Fairness**. The *Ethics Guidelines for Trustworthy AI* require the implementation of Universal Design Principles. Usability and accessibility checks are still outstanding for this service. Overall, further testing and monitoring for biases and discrimination is still outstanding to avoid biases in data set and model performance. The requirement for **Societal and Environmental Wellbeing** is not fully met yet, because the extent of the environmental impact is not fully known and mitigation measures to reduce energy use have not been implemented. The societal impact, via impact on work, is better understood and no deskilling risks are identified.

Privacy and data governance should be strengthened by empowering the users of the system further, e.g. by allowing them to flag issues. The privacy of Fediverse users whose data is made visible with the service should be considered in more detail. This topic is included on the agenda for the second EAB meeting in M24.³ The gap towards achieving **Human Agency and Oversight** can partially be explained by the fact that training for humans exercising oversight cannot yet be given, as this group is not defined yet. More safeguards for human autonomy can be considered. Finally, **Accountability** is already well on track. The overall score was lower because there is no clear mechanism for users to communicate with the developers. **Transparency** is rather well implemented already. The final gap can be accounted for with some additional information made available to the users.

Annex IV contains a table with proposed actions to improve the trustworthiness of the service further. These measures draw on best practice examples from other services, or propose new approaches, if there are no suitable best practice examples, yet.

3.3 Service 3 (T4.3): Preliminary ALTAI results

In so far as they are quantifiable, the results for service 4.3 *Media Annotation and Management (MAAM)* - which analyses the authenticity of Fediverse media assets – are presented in Figure 5 below. A short discussion of the main findings will be provided, before a list of measures will be proposed (see Annex IV – ALTAI Recommendations, T4.3 Recommendations based on M18 ALTAI).

Overall, the self-assessment (orange in diagram) underestimated the trustworthiness levels based on the implemented measures (blue in diagram). The only deviation occurred with a higher self-assessment with regards to meeting the **Privacy and Data Governance** requirement. Most improvements can be made on the requirement for **Diversity, non-discrimination, and fairness**. The *Ethics Guidelines for Trustworthy AI* require the implementation of Universal Design Principles. Usability and accessibility checks are still outstanding for this service. Overall, further testing and monitoring for biases and discrimination is still outstanding to avoid biases in data sets and model performance.

³ As stated in section 2.2, a report on this discussion can not be included in this deliverable due to the timing of this submission and the meeting.



Figure 5. Spider graph T4.3 ALTAI

The requirement for **Societal and environmental wellbeing** is not fully met, because the extent of the environmental impact is not fully known and mitigation measures to reduce energy use have not been implemented. The societal impact, via impact on work, is better understood and no deskilling risks are identified. **Accountability**. The overall score was lower because there is no clear mechanism for users to communicate with the developers. Further, The training process is not well-documented, limiting auditability of the service. **Privacy and Data Governance** should be strengthened by empowering the users of the system further, e.g. by allowing them to flag issues and withdraw consent.

The gap towards achieving **human agency and oversight** can partially be explained by the fact that training for humans exercising oversight has not yet been given to the early adopters. More safeguards for human autonomy can be considered. **Transparency** is rather well implemented already by making additional information on integrated models available to the user. The final gap can be accounted for with some additional information made available to the users e.g. on how quality of inputs (such as image compression) affects output reliability.

As for the other services, the proposed mitigation measures are included in Annex IX.

3.4 Service 4a (T6.1): Preliminary ALTAI results

In so far as they are quantifiable, the results for service 6.1A *The Interactive Coaching and Simulation Module* - which provides coaching for using LLMs responsibly - are presented in Figure 6 below. A short discussion of the main findings will be provided, before a list of measures will be proposed (see Annex IV –T6.1A Recommendations based on M18 ALTAI).

The self-assessment (orange in diagram) was overall very similar to the level of trustworthiness indicated by the implemented measures (blue in diagram). Measures adopted for **Transparency** and **Societal and Environmental Well-being** are **fully adequate** for the current state of the project.

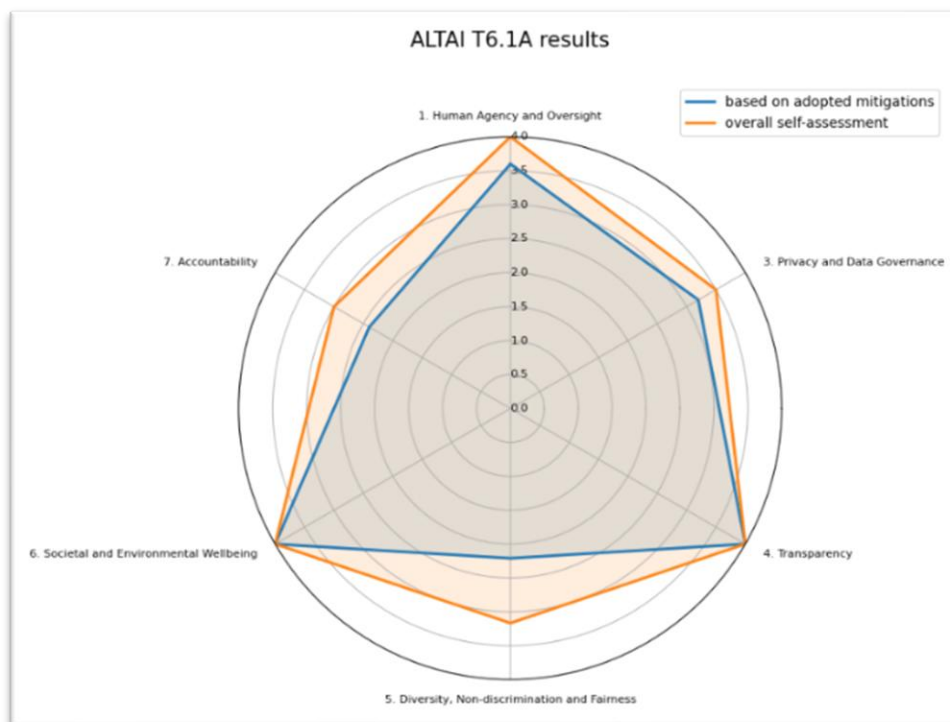


Figure 6. Spider graph T6.1A ALTAI

The greatest improvement can be made towards **Diversity, non-discrimination, and fairness**. According to the *Ethics Guidelines for Trustworthy AI*, this requires the implementation of Universal Design. Further, users don't have a way yet of flagging issues to the service developers. Understanding the actual challenges that this service poses to fairness can be improved through additional testing and monitoring.

Accountability is already rather high. Traceability of the operations can be improved, especially post-deployment. **Privacy and data governance** are facing the practical limitation of utilizing pre-trained LLMs. The AI-CODE partners therefore cannot steer data governance and privacy during training processes and lack access to understanding the underpinning data practices. Training an LLM from scratch is not a viable option. Small improvements can be made by clarifying to the user that and how they can flag concerns with regards to privacy and data protection.

Finally, the gap towards achieving **human agency and oversight** can be closed by fully training those who will exercise oversight over the AI service. Since the system is not yet widely deployed, for now this requirement is adequately fulfilled.

3.5 Service 4b (T6.1): Preliminary ALTAI results

In so far as they are quantifiable, the results for service 6.1B *The Counter-Narrative Recommendation Tool* – which provides claim verification – are presented in Figure 6 below. A short discussion of the main findings will be provided, before a list of measures will be proposed (see Annex IV – T6.1B Recommendations based on M18 ALTAI).

Overall, the self-assessment (orange in diagram) is more optimistic about meeting the Trustworthy AI requirements than the adopted mitigation measures indicate (blue in diagram). The only deviation occurred with a lower self-assessment of the **Transparency**, which has reached the full score for the M18 assessment. **Human Agency and Oversight** are also fully adequate for the current

state of the project. Most improvements can be made on the requirement for **Diversity, non-discrimination, and fairness**. The *Ethics Guidelines for Trustworthy AI* require the implementation of Universal Design Principles. Usability and accessibility checks are still outstanding for this service. Overall, further testing and monitoring for biases and discrimination is still outstanding to avoid biases in data set and model performance.

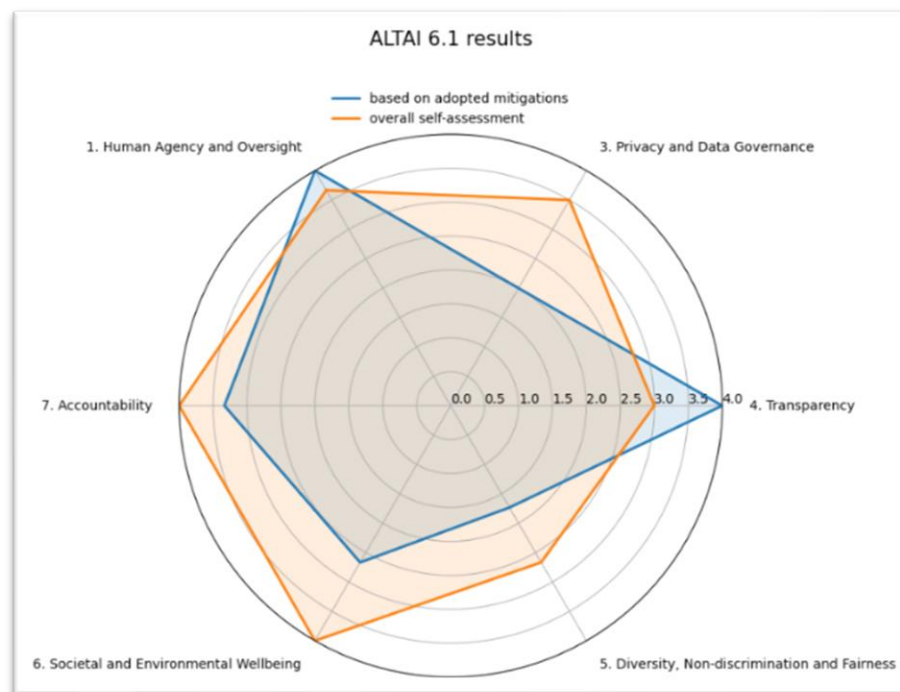


Figure 7. Spider graph T6.1B ALTAI

Privacy and data governance are constrained by the use of pre-trained LLMs. As a result, the AI-CODE partners cannot control data governance and privacy during the training phase and have limited insight into the underlying data practices. This issue is also included on the agenda for the second annual EAB meeting to receive guidance on how to best assess and mitigate the risk. Improvements can be made by clarifying to the user that and how they can flag issues with regards to privacy. The requirement for **Societal and environmental wellbeing** is not fully met, because the extent of the environmental impact is not fully known. Mitigation measures to reduce the energy use have been implemented nonetheless. No deskilling risks are identified.

Finally, **Accountability** is already well on track. The overall score was lower than it could be, due to the requirement for establishing a system for value trade-offs. However, it is not quite clear that this requirement is truly applicable to this service, because the developers are not financially dependent on the service operating in a certain way.

3.6 Service 5 (T6.2): Preliminary ALTAI results

In so far as they are quantifiable, the results for service 6.2 *Transparency Service for AI Model Cards* - which provides a standardised structure for creating and managing AI model cards - are presented in Figure 8 below. A short discussion of the main findings will be provided, before a list of measures will be proposed (see Annex IV – ALTAI Recommendations, T6.2 Recommendations based on M18 ALTAI).

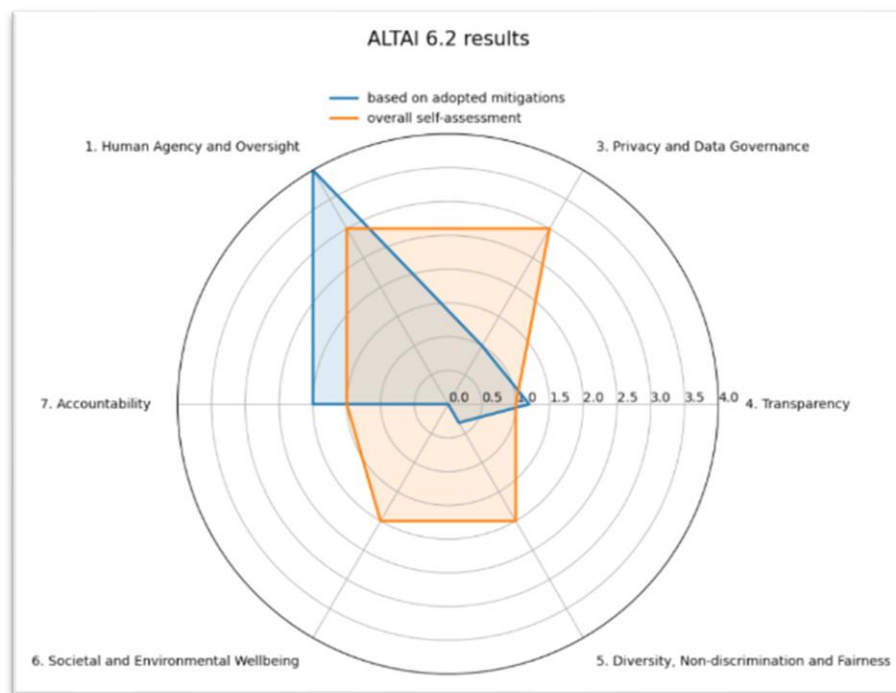


Figure 8. Spider graph T6.2 ALTAI

The self-assessment (orange in diagram) was more optimistic than the implemented measures (blue in diagram) indicate and progress towards the requirements varies greatly. The requirement for **Human Agency and Oversight** is adequately fulfilled for the current point of the service development and the low risk levels. Most improvements can be made towards the **Societal and Environmental Wellbeing**. These impacts should be addressed more comprehensively. For example, the extent of the environmental impact is not fully known for this service. There is also room to understand better how exactly this tool will impact work arrangements, considering its envisioned use in the professional setting and cognitive skills.

Diversity, Non-discrimination, and Fairness require, according to the *Ethics Guidelines for Trustworthy AI*, the implementation of Universal Design. Users don't have a way yet of flagging issues to the service developers. Overall, for this requirement the understanding of the actual challenges that this service poses can be improved through additional research and testing. However, testing LLMs in this regard is resource intensive and evaluations are difficult to design. Thus, this topic is also discussed with the EAB in their second annual meeting to receive additional guidance.

Privacy and Data Governance should be strengthened by implementing GDPR fully and empowering the users of the system further, e.g. by allowing them to flag issues. With regards to **Transparency**, the service is not yet providing users with information on the model that allows for appropriate levels of reliance and trust in the system. This can easily be improved by providing access to certain performance indicators. **Accountability** is already relatively high. The overall score was lower than it could be, due to the requirement for establishing a system for value trade-offs. However, it is not quite clear that this requirement is truly applicable to this service, because the developers are not financially dependent on the service operating in a certain way.

3.7 Service 6 (T6.3): Preliminary ALTAI results

In so far as they are quantifiable, the results for service 6.3 *Personal Companion for Understanding Disinformation* – which assists journalists and fact-checkers in combating the spread of dis- and misinformation – are presented in Figure 9 below. A short discussion of the main findings will be provided, before a list of measures will be proposed (see Annex IV – ALTAI Recommendations, T6.3 Recommendations based on M18 ALTAI).

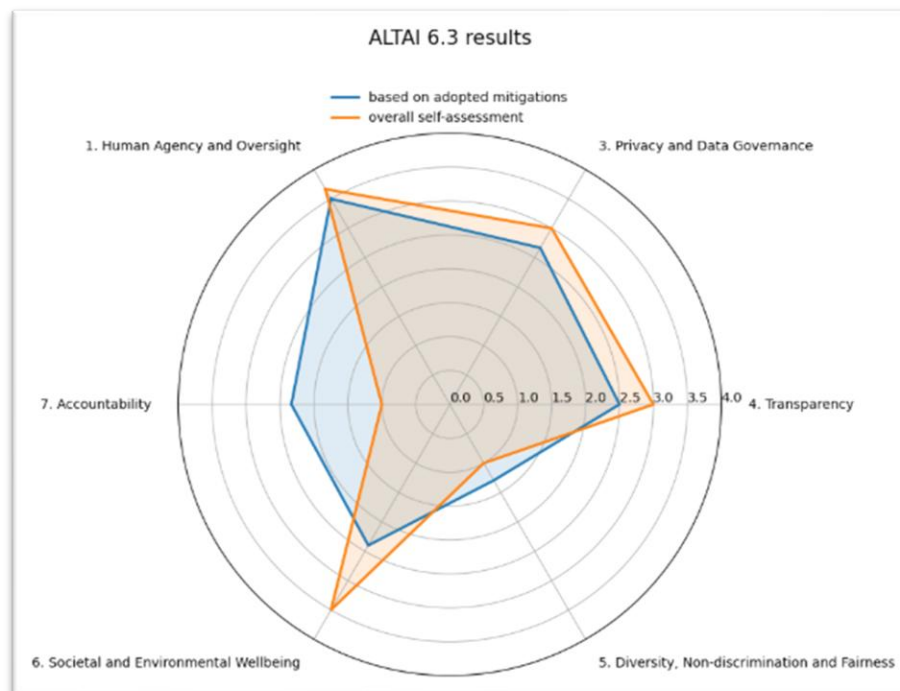


Figure 9. Spider graph T6.3 ALTAI

The self-assessment (orange in diagram) was mostly in line with the adopted measures (blue in diagram), except for an underestimation of achieved **Accountability** and overestimation of **Societal and Environmental Wellbeing**. Most improvements can be made towards **Diversity, Non-discrimination, and Fairness**, which require, according to the *Ethics Guidelines for Trustworthy AI*, the implementation of Universal Design and extensive testing. Users don't have a way yet of flagging issues to the service developers. **Accountability** is already relatively high and almost fully adequate for the current state of the project. The overall score was lower than it could be, due to the requirement for establishing a system for value trade-offs. However, it is not quite clear that this requirement is truly applicable to this service, because the developers are not financially dependent on the service operating in a certain way.

The impact on **Societal and Environmental Wellbeing** should be addressed more comprehensively. For example, the extent of the environmental impact is not fully known for this service. Some mitigation measures have been adopted, nonetheless. Materials for re- and up-skilling measures can help mitigate potential negative societal impacts. The gap towards achieving **Transparency** can be explained by the fact that accuracy and error rates will be added to the information accessible to users once significant testing has been done. There are currently also no input quality controls in place, which has lower priority as this is not a continuous learning system.

Privacy and Data Governance should be strengthened by empowering the users of the system further, e.g. by allowing them to flag issues. A practical limitation for reaching full trustworthiness in this criterion lies in the use of a pre-trained LLM. Training an LLM from scratch is not a viable

option. Consequently, the AI-CODE partners cannot shape data governance and privacy during training and have limited understanding of the data practices on which the model relies. Due to the complicated nature of this issue, it was brought up at the second annual EAB meeting. Finally, the gap towards achieving **Human Agency and Oversight** should be seen as adequately fulfilled at this point in the project, as risk levels are low.

4. Methodological Reflections

This report detailed the specification and implementation of the EbD approach in the project under T9.3. As evident in Figure 1 and the description in this deliverable, it spans a wide range of activities that were conducted in parallel and with the involvement of the entire consortium. Such an ambitious approach is not without challenges, and EbD is not without critics, as presented in section 1. In this part of the report, we take a step back to reflect on advantages and challenges of the chosen approach.

4.1 Success of EbD in AI-CODE

Overall, the advantage of EbD as we are implementing it in AI-CODE has been the collaboration it has necessitated across consortium members. Service developers in particular have been involved in all the conducted activities. This allowed the activities related to ethics to be specific and relevant to the AI-CODE project. Further, the joint activities have contributed to a sharpened awareness of ethical issues in the technologies we design and build and an increased understanding of the importance and relevance of ethics for the day-to-day activities of AI development. As research suggests, AI developers are often not fully aware of the extent to which they are making ethical decisions, and thereby not cognizant of their moral agency (Griffin et al., 2024). By involving the AI developers in EbD activities, as well as utilising ALTAI as an educational tool, we are raising the awareness of ethical issues in the development of AI.

Another main advantage was a learning experience for the ethical experts at CEPS. With the close focus on the AI-CODE services, it has been necessary to specify the ethical problem for the specific project context already. Thereby, the specification problem of applied ethics in applying abstract ethical principles (Beauchamp, 2003) was counteracted by the service-specific framing of all of our activities and analyses. The nature of conducting EbD on AI tools as they are under development also necessitated a certain pragmatism in our criticism. Rather than offering broad, sweeping criticism, we have been searching for mitigation measures that allow for better ethical alignment with Trustworthy AI in lockstep with identifying issues.

With the emphasis on proposing mitigation measures, such as those agreed upon after the first annual EAB meeting and based on the ALTAI assessment, we are strengthening the ethical practices in a meaningful way. Meanwhile, this practice-oriented collaboration can also support the moral agency potential of service developers, who might add some of the measures implemented on AI-CODE to their mental toolbox for ethical AI development. By centring dialogue-based collaboration across disciplines we build trust between parties. Practising ethics in this way is not only about prohibiting some things or creating obstacles for research and innovation. Instead, it becomes a proactive and shared responsibility between actors, which come together to foster the shared goal of developing good – both morally and technically – services.

4.2 Challenges of EbD implementation

While those advantages of the chosen ethical approach have manifested in the AI-CODE project, some challenges have also emerged. The main challenge that has presented itself over the first two years has been the increased burden for service developers. While they are working on the technical development and implementation of the AI-CODE action, the EbD approach has added additional requirements for the technical development. The demands on developers from EbD are particularly complex because they require the translation of ethical principles into technical measures and thinking across disciplines. Furthermore, the AI-CODE service developers were not only asked to attend to ethical concerns, but also to the particular feedback from their user groups, i.e. media professionals, fact-checkers, journalists and journalism students.

Multiple actions were taken to ease the burden and time-intensity of the EbD activities on the service developers. For example, the M18 ALTAI assessment was implemented in M20/21 after the deliverables of WP4 and 6 have been submitted successfully. Further, the EAB meetings are facilitated by CEPS and service developers are not expected to attend. Instead, they can rely on the timely transmission of the important outcomes. Fortunately, the involvement and feedback from users also contributes to the EbD framework. Accounting for the epistemic faculties and perception of users matters for Trustworthy AI and can be recognised as a synergy between tasks involving users and those on ethics.

Another challenge is that the chosen ethical framework for AI-CODE – *The Ethics Guidelines for Trustworthy AI* – consists of general guidelines for AI ethics. The principles-based approach chosen by the AI-HLEG can also be criticised for neglecting more traditional perspectives of applied ethics, such as analyses with theories such as liberalism or egalitarianism. Many researchers have published works to fill this gap, such as *Toward a Theory of Justice for Artificial Intelligence* (Gabriel, 2022) or *The Political Philosophy of AI* (Coeckelbergh, 2022).

The goal of the ethics approach chosen in AI-CODE was to practically guide the specific implementation of ethics in AI-CODE and not the philosophical exploration of speculative issues. The *Ethics Guidelines for Trustworthy AI* were chosen for this goal because of their breadth, practical usefulness, impact and expert consensus. While this might be a legitimate critique of the *Ethics Guidelines for Trustworthy AI*, namely that they do not invite the more traditional ethical discussion which employs moral theories, it is not a critique that makes the guidelines ill-suited for the objectives of T9.3.

A more relevant critique of the chosen ethical framework is that it does not account for the specific ethical context of journalism. While professional ethics are considered through the participation and involvement of journalists throughout the co-creation activities, in the final year of the project, they should be spotlighted further in the activities of T9.3 and in the Deliverable *D9.6 Ethics Handbook* (M36) in particular. As discussed in section 1.1, such context-specific assessments should account for ethical, legal and social aspects (ELSA) (van Hilten et al., 2025). For this purpose, the M24-36 activities will focus more concretely on ethical, legal, social and economic (ELSE) aspects. In expanding the scope of consideration to all ELSE aspects, we can expect both synergies with ethical issues already addressed but also possibly conflicting demands, e.g. based on the economic or social aspects. Thus, even when enriching EbD with context-sensitive ELSE frameworks, achieving ethical AI will remain challenging. Finally, a specific limitation to our EbD approach has become apparent in the use of ALTAI.

4.3 ALTAI: Limitations and improvements

Whilst conducting the M18 interviews with ALTAI, we were faced with a challenge of adequately answering some questions. This challenge was related to the fact that ALTAI predates generative AI (GenAI). Gen AI models, like large language models (LLMs), increasingly form an essential part of other AI applications. Gen AI models are usually general-purpose AI (GPAI), which means that the models can be used for varying different purposes. Through a process of finetuning GPAI models on context-specific data, they can be adapted to different use cases and deployment contexts. This way, GPAI models become used in varying different downstream applications. While this allows developers to build on the complex functionalities of pre-trained models rather than endeavouring in the resource- and expertise-intensive process of developing one from scratch, it raises ethical concerns because the developers of downstream applications don't oversee major parts of the development process. This becomes particularly challenging to tackle because most GPAI models are developed under proprietary circumstances.

For example, one issue around LLMs is the lack of transparency on concrete development practices⁴. Questions that are relevant for accountability, such as ensuring traceability of the development process and the sourcing of training data are beyond the knowledge and control of the downstream deployer for the GPAI model. Of course, the fine-tuning process can be documented, but it is impossible for downstream developers to ensure traceability as it is included in the ALTAI. Further issues arise with respect to the principle of privacy and data protection. ALTAI contains the question: "Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle?" Due to the lack of openness on the data used, we do not know for many LLMs whether privacy was respected for the full lifecycle. Research suggests⁵, though, that Meta – the developer of the Llama models – has not followed privacy protection throughout. 'Privacy' is not only an ethical requirement for trustworthy AI, it is a core principle for journalism ethics, as well. If a journalist were to report something using AI-CODE tools based on private data revealed from the training data, this would be a severe breach of professional standards.

The non-personal data protection is also important when questions of copyright arise. Multiple news outlets, such as the New York Times, have sued AI companies over copyright breaches. If AI-CODE's services rely on LLMs that undermine professional standards (e.g. privacy and transparency) and business models, this seems to lead to an ethical trade-off between the potential usefulness of the tools and their impact on the profession overall. With the lack of transparency on training data, concerns arise with regards to accuracy and reliability in downstream applications. The datasets used to train models must be broad and diverse enough to enable models to accurately deal with edge cases of hate speech or misinformation to be reliable for the AI-CODE services. Downstream users of GPAI models rarely have such detailed insight into the training data set to answer this question. Further, for ALTAI questions relating to the accuracy of AI systems, it is difficult to assess the accuracy of natural language outputs. There is always a range of answers that are possible and factually correct but that can have different connotations.

Another issue relates to meeting the requirement for 'Diversity, Non-Discrimination, and Fairness' in ALTAI, which builds on defining fairness within an algorithmic context. Since LLMs use natural language, applying a fairness definition is not a straightforward exercise. Further, bias cannot be

⁴ <https://crfm.stanford.edu/fmti/May-2024/index.html>

⁵ <https://www.private-ai.com/en/blog/privacy-llama2-gpt4-gemini>

removed fully from natural language. For example, Shannon Vallor points out: “The model does not independently acquire racist or sexist associations. We feed these associations to the model in the form of training data from our own biased human judgements.” (Vallor, 2024, p. 43) It is unclear how to interpret the requirements on fairness for LLM outputs and how to realistically evaluate or mitigate them.

Given that the majority of user-driven services in AI-CODE build on generative AIs, particularly the Llama models, some ALTAI questions could not be answered in a satisfying manner. For GPAI models, much of the AI life cycle occurs before and outside of the control of the AI-CODE service developers. These are issues related to LLMs that are carried downstream. We cannot solve these in the AI-CODE project, but we aim to use LLMs as ethically as possible within our sphere of influence. Since ALTAI does not provide sufficient guidance in this regard, we are consulting with the EAB in the second annual meeting (M24) on how to best deal with the downstream issues of LLMs that are not fully within our control. For example, we need to determine how much testing is realistic and needed from the AI-CODE consortium to consider its services trustworthy.

We intend to share this knowledge with the consortium and in the next public report Deliverable D9.6 *Ethics handbook* and suggest an update of ALTAI in light of the challenges posed by generative AI by producing an academic policy-oriented paper.

5. Conclusion

This report demonstrates that the ethical dimensions of the AI-CODE services were taken into account throughout the entire design and development process by employing an EbD approach. To this end, we started by identifying and mapping ethical issues and areas of concern, enabling a comprehensive assessment and prioritisation of issues for mitigation and monitoring. The contributions from the EAB, which meets annually, further strengthened this approach. In M12 of the project, the EAB provided feedback for each service, which enabled important adjustments to the design of AI-CODE services in order to better address ethical concerns.

The most important step in pursuing an EbD approach was the selection of an ethical framework to guide our efforts. The *Ethics Guidelines for Trustworthy AI* were selected due to their importance, impact, breadth, and practice orientation. They specify seven principles for trustworthy AI, which were translated into more practical high-level design requirements for AI-CODE services. The implementation of these design requirements has been monitored through an ethical evaluation based on the accompanying *Assessment List for Trustworthy AI* (ALTAI). The ALTAI has been adapted to the needs of the project, as the AI-HLEG recommends in their introduction to this assessment tool. This adaptation involved asking 91 ALTAI evaluation questions in a first round, whilst postponing the evaluation of technical robustness and safety until higher technological readiness levels are achieved. We assigned some questions to users, since they are better positioned to answer whether the services are sufficiently transparent to allow users to understand their functioning. The preliminary results of the assessment at the halfway point of the project were overall positive and raised awareness of additional steps that can be taken. Together with the recommendations from the EAB in M12, the AI-CODE services have received concrete suggestions on how to improve the ethical dimensions of their tools.

Overall, this ambitious EbD approach has offered several advantages, including interdisciplinary collaboration, increased awareness of ethical issues in AI development, and an orientation towards

finding solutions. However, we also identified three main challenges over the course of the first two years. First, the EbD activities have added considerable workload to AI-CODE service developers, requiring interdisciplinary translation of expectations as well as collaboration on identifying feasible mitigation measures. Second, the ethical evaluation tool, ALTAI, is partially outdated given the recent advancement in AI technologies, especially regarding GenAI. Finally, we recognise a lack of specific contextual sensitivity in the *Trustworthy AI* framework for the context of journalism. The final two shortcomings will be addressed over the course of the next year of the project.

The AI-CODE ethics-related work will continue throughout the third and final year of the project. By the end of M36, we will conduct an exhaustive assessment of the ethical, legal, social, and economic (ELSE) aspects of AI-CODE. This will result in a comprehensive *Ethics Handbook* (D9.6), which includes results from the project’s ethics monitoring, best practices and ethical standards. This ELSE assessment allows us to evaluate AI-CODE tools in the specific context of future deployment, i.e., journalism and fact-checking. Thereby, our ELSE assessment will take into account the ethical, social, and economic complexity of the media sector and elaborate on the kind of adjustments required to allow for a positive impact of the AI-CODE services.

During this upcoming period, we will also complement the ethical assessment by conducting a second ALTAI round, which will include the *Technical robustness and safety* requirement, as well as all the questions postponed for higher technical readiness levels. The shortcomings of ALTAI in light of advancements in GenAI have been discussed with the EAB in M24. Their comments will allow us to adjust outdated ALTAI questions to better illuminate how AI-CODE services can be trustworthy while using pre-trained LLMs. And, of course, we will host the third and final EAB meeting to discuss any final loose ends.

In summary, the upcoming year is critical for the ethical design of the AI-CODE services for three main reasons:

- It will allow the project to validate whether the ethical design requirements identified so far have been effectively integrated into the services.
- It will address ethical gaps that could not be resolved earlier – especially those related to generative AI, user-facing transparency, and Fediverse data considerations – by conducting the second ALTAI assessment in M36, the ELSE analysis, as well as further EAB consultations in M24 and M36.
- With the UI prototypes and user testing scheduled in the final year, the AI-CODE services will be able to implement, test, and refine the concrete ethical safeguards that the earlier work has defined.

Overall, the goal of trustworthy AI-CODE services has guided all project activities. Ethics is not an afterthought; it is actively addressed through the collaborative efforts of the entire consortium.

6. References

- Allein, L., Moens, M.-F., Perrotta. (2023). Preventing profiling for ethical fake news detection. *Information Processing & Management*, Volume 60, Issue 2. <https://doi.org/10.1016/j.ipm.2022.103206>.
- Beauchamp, T. L. (2003). The Nature of Applied Ethics. In *A companion to applied ethics* (pp. 1–16). Blackwell.
- Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics* (6. ed). Oxford Univ. Press.
- Bontridder, N., & Pouillet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3(3), 1-21. Article e32. <https://doi.org/10.1017/dap.2021.20>
- Brey, P., Dainow, B. (2024). Ethics by design for artificial intelligence. *AI Ethics* 4, 1265–1277. <https://doi.org/10.1007/s43681-023-00330-4>
- Buolamwini, J., Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81:1–15. Conference on Fairness, Accountability, and Transparency
- Carillon, K. (2024). "An Algorithm for Public Service Media?" *Embedding Public Service Values in the News Recommender System on RTBF's Platform*. *Emerging Media*, 2(3), 422-448. <https://doi.org/10.1177/27523543241290976> (Original work published 2024)
- Cartwright, Barry; Frank, Richard; Weir, George; Padda, Karmvir; and Strange, Sarah-May. (2023). "Deploying Artificial Intelligence to Combat Covid-19 Misinformation on Social Media: Technological and Ethical Considerations". *Hawaii International Conference on System Sciences 2023 (HICSS-56)*. 5. https://aisel.aisnet.org/hicss-56/dsm/data_analytics/5
- Choudhury, A., & Shamszare, H. (2023). Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research*, 25, e47184. <https://doi.org/10.2196/47184>
- Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.
- Coeckelbergh, M. (2022). *The political philosophy of AI: An introduction*. Polity press.
- Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & De Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10), 100857. <https://doi.org/10.1016/j.patter.2023.100857>
- Deuze, M. (2005). What is journalism?: Professional identity and ideology of journalists reconsidered. *Journalism*, 6(4), 442–464. <https://doi.org/10.1177/1464884905056815>
- European Commission. (2019). *Ethics guidelines for trustworthy AI*. Publications Office.
- European Commission. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. European Commission.

- European Commission. (2021). *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*.
- Gabriel, I. (2020). *Artificial Intelligence, Values, and Alignment*. *Minds & Machines* 30, 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Gabriel, I. (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, 151(2), 218–231. https://doi.org/10.1162/daed_a_01911
- Gillespie, T. (2023). The Fact of Content Moderation; Or, Let’s Not Solve the Platforms’ Problems for Them. *Media and Communication*, 11(2), 406–409. <https://doi.org/10.17645/mac.v11i2.6610>
- Griffin, T. A., Green, B. P., & Welie, J. V. M. (2024). The ethical agency of AI developers. *AI and Ethics*, 4(2), 179–188. <https://doi.org/10.1007/s43681-022-00256-3>
- Johnson, D. G., & Verdicchio, M. (2025). The sociotechnical entanglement of AI and values. *AI & SOCIETY*, 40(1), 67–76. <https://doi.org/10.1007/s00146-023-01852-5>
- Juhasz, C., Gianni, R., Kamalski, J., & Meacham, D. (2025). Navigating many voices: lessons for ELSA/RRR from the pragmatist perspective of ELSA-by-design. *Journal of Responsible Innovation*, 12(1). <https://doi.org/10.1080/23299460.2024.2445308>
- Kymlicka, W. (2002). *Contemporary political philosophy: An introduction* (2nd ed). Oxford University Press.
- Lisker, M., & Mihaljević, H. (2025). *Data Ethics in the Fediverse: Analyzing the Role of Instance Policies in Mastodon Research* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2505.07606>
- Morley, J., Elhalal, A., Garcia, F. et al. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds & Machines* 31, 239–256. <https://doi.org/10.1007/s11023-021-09563-w>
- Morley, J., Kinsey, L., Elhalal, A. et al. (2023). Operationalising AI ethics: barriers, enablers and next steps. *AI & Soc* 38, 411–423. <https://doi.org/10.1007/s00146-021-01308-8>
- Noble, S. U. (2019). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.
- Nissenbaum, H. (1996). Accountability in a computerized society. *Sci Eng Ethics* 2, 25–42. <https://doi.org/10.1007/BF02639315>
- Prates, M.O.R., Avelar, P.H. & Lamb, L.C. (2020). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Comput & Applic* 32, 6363–6381. <https://doi.org/10.1007/s00521-019-04144-6>
- Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, 26(5), 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>
- Tännsjö, T. (2013). *Understanding ethics* (Third edition). Edinburgh University Press.
- Vallor, S. (2024). *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*.

Oxford university press.

- van den Hoven, J. (2007). ICT and Value Sensitive Design. IFIP International Federation for Information Processing, Volume 233. The Information Society: Innovations, Legitimacy, Ethics and Democracy, eds. P. Goujon, Lavelle, S., Duquenoy, P., Kimppa, K., Laurent, V., (Boston: Springer), pp. 67-72.
- van Hilten, M., Ryan, M., Blok, V. and de Roo., N. (2025). Ethical, Legal and Social Aspects (ELSA) for AI: An assessment tool for Agri-food. Smart Agricultural Technology, Volume 10.
<https://doi.org/10.1016/j.atech.2024.100710>.
- Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20.
<https://doi.org/10.1145/3544548.3581318>
- Zwart, H., Nelis, A. (2009). What is ELSA genomics? Science & Society Series on Convergence Research. EMBO Rep. 10(6):540–544. doi: 10.1038/embor.2009.115

Annex I – Mapping of ethical areas of concern

The complete board is available here:

https://miro.com/welcomeonboard/VFd3VWVaMzYzUnRNTHZoaXlvSURqMG05MC9TYINVNlo5NS9HUG0vWHpyL3VCQ3ZMb0t3bC9DamIOUzZEQUYrWEJmZEFwU2dvMVRUbmVDU1BaNUFPeUIEOWY4MzNhLOZBaFdjREpHMGwwVWlOM2htaW1DQTY5aWsxRU9zSWw0dGRzVXVvMm53MW9OWFg5bkJoVXZxdFhRPT0hdjE=?share_link_id=823669503535

Screenshots of each service specific mapping are included below,

Mapping for T4.2 Trustability and credibility assessment

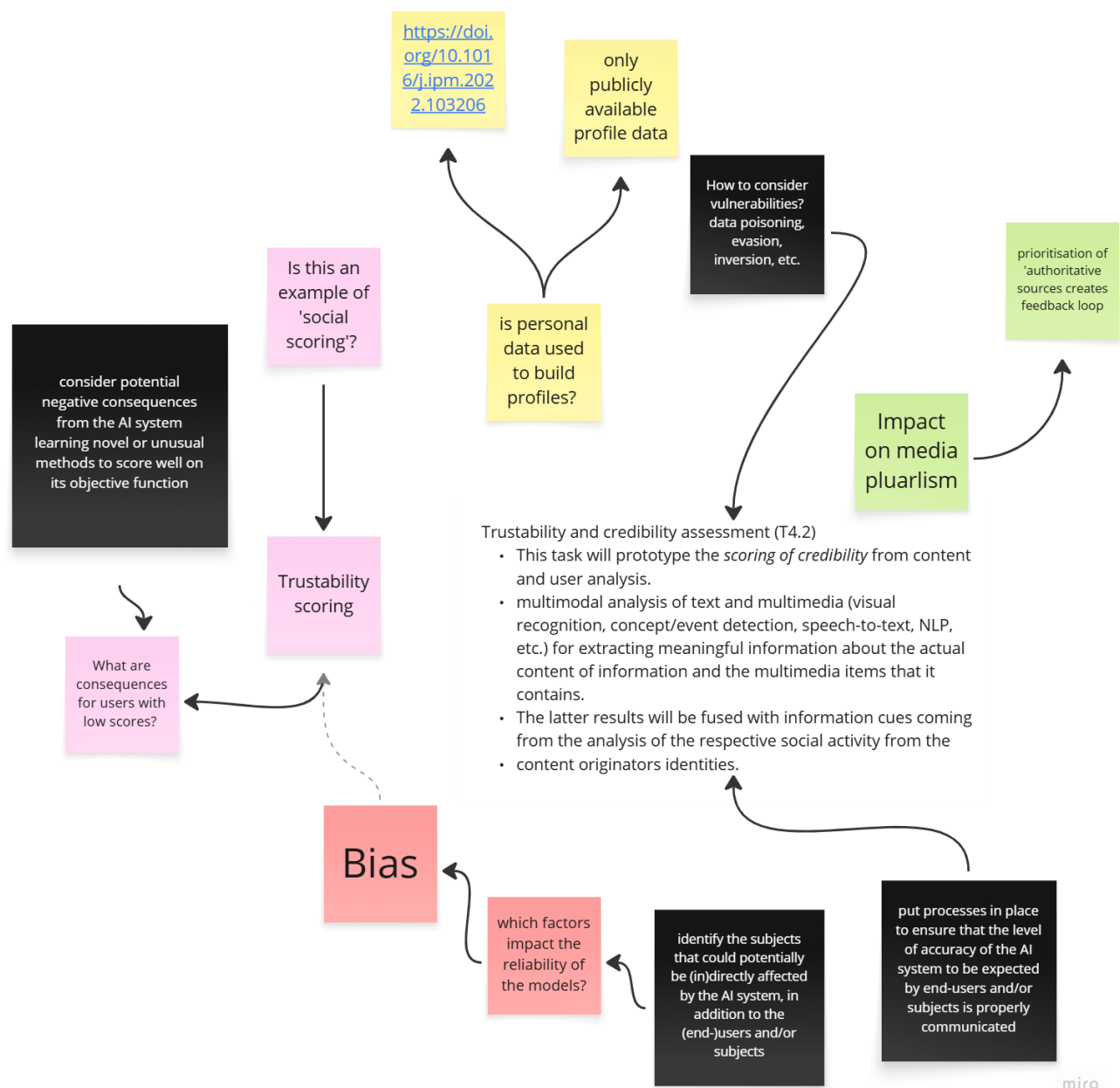


Figure 10. Mapping of ethical concerns for T4.2

Mapping for T4.3 (MAAM)

Media Asset Annotation and Management (MAAM) (T4.3)

- develop and improve MAAM and its capabilities in the context of discovering and assessing multimedia content.
- act as a multimedia repository for the project use cases and provide advanced features, such as comprehensive user profiles, including expertise, history of contributions, and credentials to enable end-users to evaluate the trustworthiness of other users.
- MAAM will be enriched with features enabling media assets collection from diverse sources, such as TikTok.

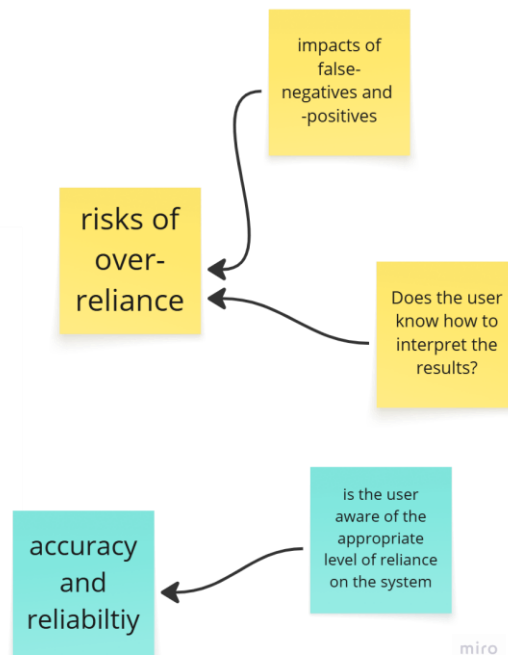


Figure 11. Mapping of ethical concerns for T4.3

Mapping for T6.1 Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals

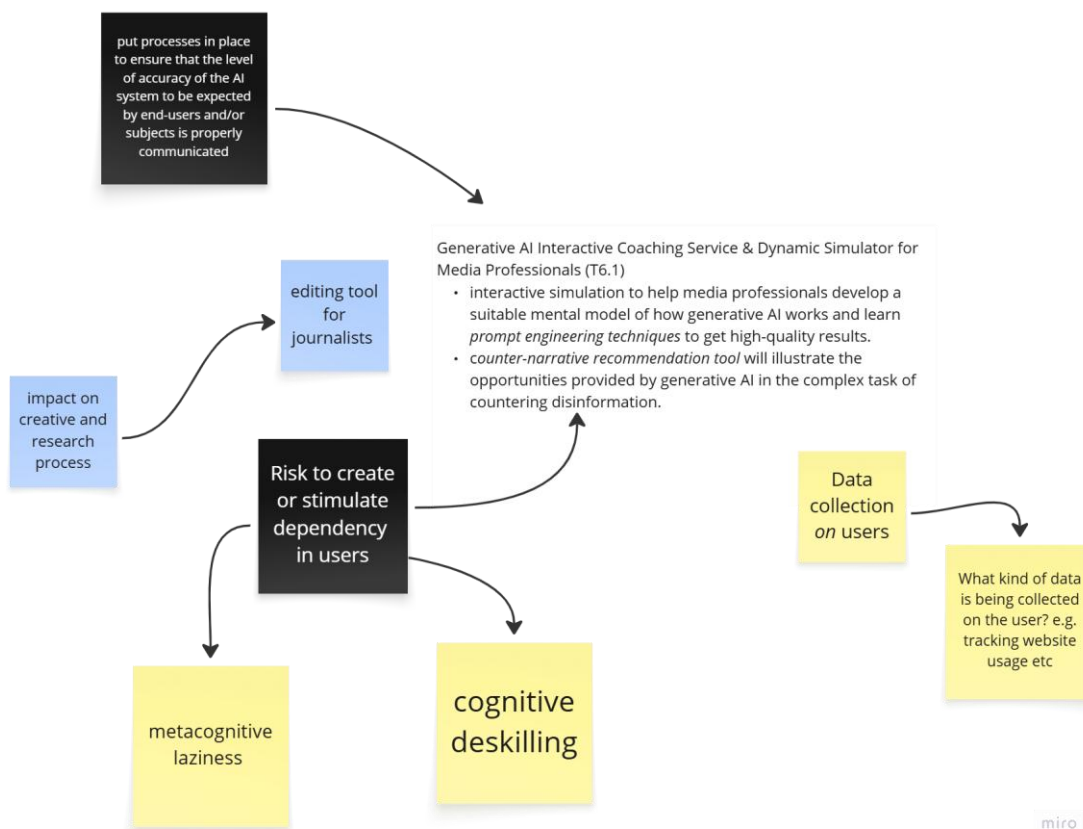


Figure 12. Mapping of ethical concerns for T6.1

Mapping for T6.2 Transparency Services for AI-CODE

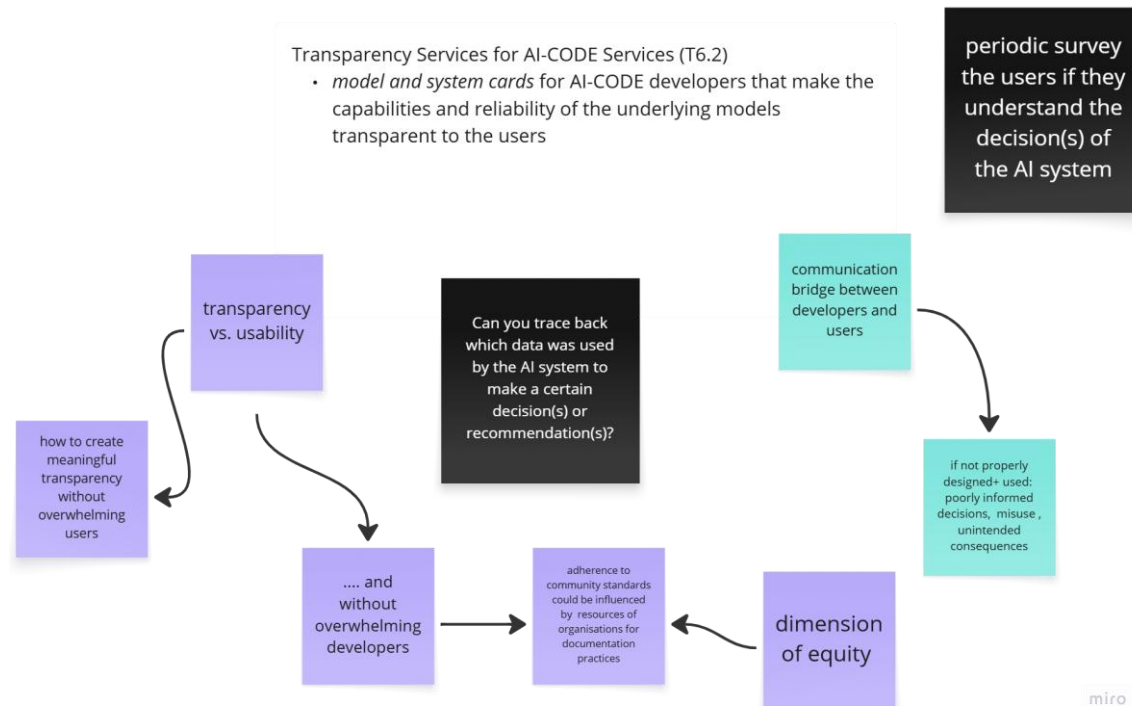


Figure 13. Mapping of ethical concerns for T6.2

Mapping for T6.3 Personal Companion for understanding disinformation for media professionals

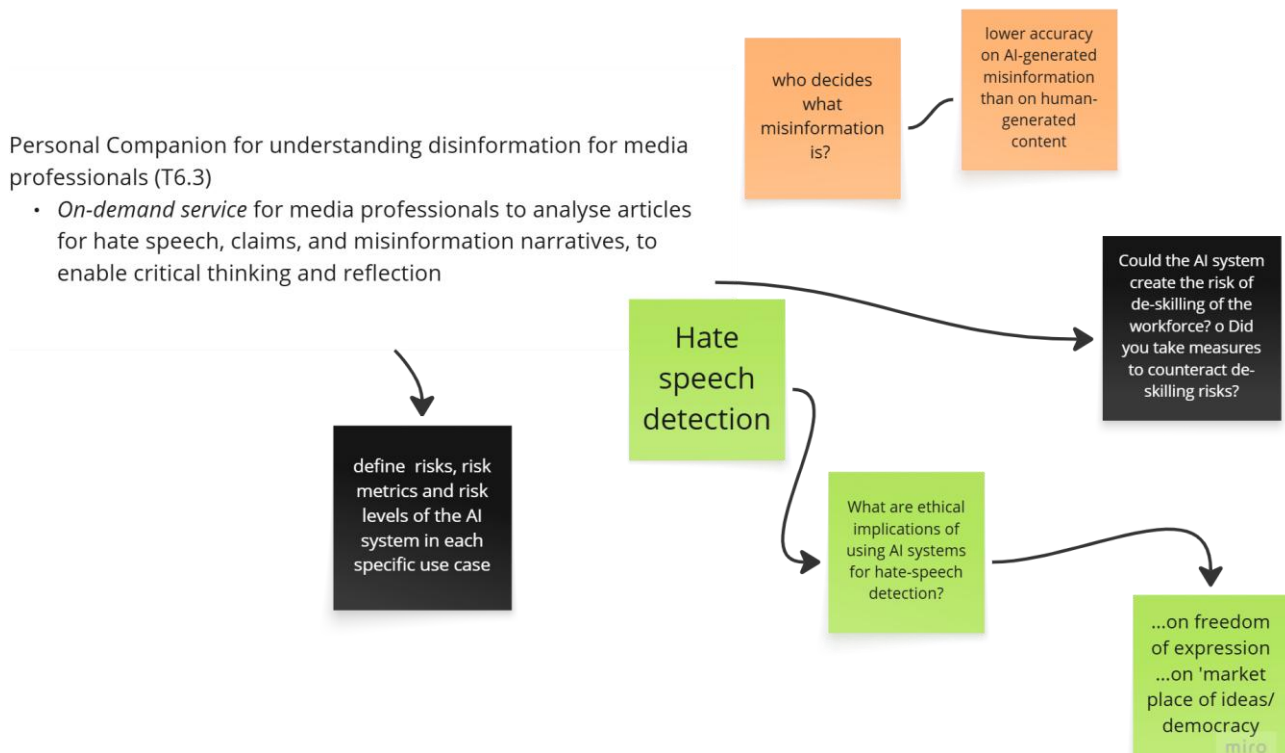


Figure 14. Mapping of ethical concerns for T6.3

Annex II – Measures based on EAB Feedback

As mentioned in section 2.2, the EAB meeting was documented by CEPS in a minutes document, and the recommendations and clarification questions of the EAB were discussed with each service developer team to decide on implementation measures. These were documented in a table which was uploaded to the project SharePoint. They are included here in Annex II for each service.

T4.1 Disinformation Detection in the next generation social media

Comments/ clarification

Do the service developers review the information regarding why Instances are banned or which keywords have been assigned to Instances (e.g. hate speech, racism, etc.)?

- No. Indeed, community-organised banning is a documented phenomenon. There are methodological challenges that make such a review difficult to implement. For example, misinformation might not be shared every day on an instance, so that the network moderation seems unwarranted on one day, while it might be accurate overall. Instance keywording is crowd sourced, so there needs to be a critical mass of classifications by multiple administrators. There are also public block lists that are curated by the list creator. In the table below some mitigation measures are proposed to enhance transparency about this process.
- Clarified: it's a tool for exploring previous network moderation decisions. 'Classification' of instances is not taking place.

The EAB members would have liked to receive more clarity on the risks/ performance limitations: How are keywords for Instances assigned? How are the indexed keywords decided on and who assigns them? Are they only indexed in English? Does this align with the demographics of the user base? Are users using high English or using colloquial language? What kind of words are indexed and in which language? Which languages are being used and how is this accounted for in the service? Is the service already taking potential biases of classifications into account? How have we tuned our services for these?

- The network moderation decisions usually use English terminology, even on non-English speaking instances the same keywords are typically used. The bias towards English does not depend on this AI-CODE service, but on the crowdsourced information. In the sense that the network moderation keywords are human-dependent, it's biased. The exploration tool does not pre-select instances or network moderation events. All instances that are online in the moment of checking are included. That means that there is no selection bias because there is no selection.

Are most prominent user accounts shown? It could easily be made available in the inspector tool, but the team has rejected the idea.

The EAB also discussed the question posed by the consortium on using user information, which is publicly available, in the context of AI-CODE research. The EAB agreed that this question is worthwhile to address, because the data is not anonymised (the service identifies specific users). It appears permissible to process this data, however, we should not publish any names of users, because that is personal information.

Table 2. EAB recommendations for T4.1

EAB recommendation	Implementation measure/ Clarification
Provide information on risk factors for misclassification of instances, to avoid a ‘halo effect’ and over-reliance. (For example, an Instance might seem overall credible based on the analysis of this service. This might influence the perception of the service user that all content coming from it is reliable.)	- The service will provide additional quantified measures, e.g. this instance has been evaluated as ‘hate speech’ by X other instances - It will be considered if data on network moderation history (instance classifications change over time – keywording is a moving target) should be included. - A small disclaimer will be implemented that notifies the user that the service relies on crowdsourced information
Review the correct classification of instances	Due to the reasons above, this was considered not to be feasible. It was clarified that this network moderation tool is only a starting point. The inspector tool (T4.2) can be used for more substantial analysis. This task division will be made more apparent. A small disclaimer will be implemented that notifies the user that the service relies on crowdsourced information or making this more apparent in the name will be considered.
Clarify who has responsibility for the Instance labelling and who can be held accountable for a misclassification of Instances as unreliable and vice versa	Responsibility for mis-labelled instances does not lie with service providers, as it is social, crowd-sourced information. This tool is not changing or selecting this data, it is just aggregating and reporting. This will be made transparent to the user through the previously mentioned disclaimer.
If at all, the EAB recommends to identify and showcase the most prominent post and user of the last three months (or a timeframe similar to this) rather than an ‘absolute’	User names are not included and will not be considered without social scientists in the loop. A restriction of the time frame is already considered.
Particularly in publications, the account name of the most popular users should not be included to prevent negative consequences for users.	Account names are not included in the current version of this tool, or current publications.
Reliability of the integrated modules should be communicated to the users. // Overall: the access to information and transparency should be enhanced	As there are no ML models used in this service, but only grounded indicators, model cards don’t seem to be appropriate. Instead, method section (similar to: https://covid19obs.fbk.eu/#/methods) will be included.

T4.2 Trustability and Credibility Assessment

Comments/Clarification

Sensitivity for language used in communication with end users matters a lot.

- True/false wording should be avoided, referring to high vs. low credibility content instead
- The EAB further pointed out that there are two things in the service: i) Is the content artificial? ii) Are the propagation and language patterns (e.g. hate speech detection) indicative of credible content?

Which modules are integrated in this service? Bot identifier and the propagation graph analysis.

- The claim identifier shown in the mock-up is not integrated yet.
- The post image analysis is not yet integrated either. Could be a future collaboration with CERTH on their deep fake detection model.

Is this service disclosing the most popular/prominent originator account names?

- This type of information on accounts is not shown by this service. The service will disclose how often a search for the specific URL has been performed.
- Propagation graph: does not disclose account names, either. Example in D2.1.
- In research publications: no account names will be published. They might include numerical results on model performance and dummy data for testing, no actual data.

Table 3. EAB recommendations for T4.2

EAB recommendation	Implementation measure/ Clarification
It should be acknowledged that there is a difference between fact and claim.	Not yet relevant – possibly when other modules are integrated.
Create more transparency around the claim identifier, its parameters of operations, underlying definition, and limitations.	Not yet relevant.
Ensure that the user has enough information on performance levels of the bot identifier for the right level of reliance.	The metrics of evaluation, what they mean, and what that implies for reliability can be declared in the UI. This will be considered. Model cards will be implemented as an additional feature.
Reliability of the integrated modules should be communicated to the users.	Model cards will be implemented as an additional feature.
Enhance transparency by providing an ‘artificiality	‘Artificiality score’ can be used for those models that investigate synthetic media

score’ and a ‘credibility score’ instead of using one score.	content. ‘Credibility score’ can be used for the other metrics that are being developed, e.g. propagation graph analysis. UPM is working on an integrated score.
--	--

T4.3 Media Annotation and Assessment Management (MAAM)

Comments/ clarification

What happens to data that users upload for MAAM for analysis?

- Media assets are analysed by the models, but no fine-tuning is undertaken based on the data provided by the users.

In the description of the service, it says: “User profiles will be enriched to capture users' expertise in disinformation analysis, making the platform even more relevant for experts in the field.”

- The project proposal is deliberately open-ended as of now. The general idea is to first allow users to opt into a system where they can contribute with their expertise to improving the platform. Users should be able to provide feedback on different models that are integrated into the MAAM, such as the accuracy or false-positive/ false-negative predictions. MAAM users that are not expert users – e.g. not part of a fact-checking organisation – can find an expert-user inside MAAM to collaborate with on the assessment of a specific (potentially synthetic) asset.
- Each user has a private workspace by default, but they can also open up collaborative workspaces where users can invite people to join the workspace and have access to the shared assets. Malicious users cannot join without invitation.
- The team is currently working on creating public workspaces. In that case, users should be made aware that all MAAM users can see all of their uploaded assets.

Table 4. EAB recommendations for T4.3

EAB recommendation	Implementation measure/ Clarification
Inform users about the terms of services Clarify for potential users: • Where is uploaded data stored? • Is it used for further training of the system? • Can the user delete the images permanently? • Who can see the uploaded assets?	A pop-up window will be implemented when starting the service that informs the end user about the terms of service tailored to the specific workspace (private/shared/public), that they are not using data for further training of the system, that assets can be deleted permanently, that visibility depends on selected options.

The user should have an option to permanently delete content they upload.	MAAM platform already offers this option: Content, metadata, and predictions for assets can be fully deleted, as well as user feedback about the asset analysis. If feedback on assets shall be handled differently in the future, an ethics consultation will take place.
Enhance explicability of system outputs by creating more transparency on performance of the models. There are AI-CODE models for synthetic-images detection or forensics analysis, and models that are in place for retrieval, annotation, person recognition etc. in the backend of the service.	The user is often unaware of the backend systems, e.g. people-recognition. Additional information will be provided on these models, so that interested users can also understand e.g. embedded biases. The prediction models are more visible to the end-user. Model cards will be implemented for each of these models, building on the already implemented functionality of accessing information on the model.
Make the user aware of potential pitfalls and weaknesses in the system, such as varying levels of accuracy, to ensure appropriate level of reliance on the system.	Model cards will be implemented that include details on the limitations.
Avoid profiling if user profiles are analysed to improve the platform.	For now, there is no intention for user profiling. This will be monitored as decisions are being made for further improvements of the platform.

T6.1 Generative AI Interactive Coaching and Simulation & Counter Narrative Recommendation Service

Comments/clarification

Module 1: Perceived to be useful to train people for critical thinking with regards to sources, understanding blind spots, biases and hallucinations of generative AI

Module 2: The EAB members also asked about the sources that a researcher can upload:

- Can the user enter multiple sources? What kind of sources can be provided?
- Can the service also track where sources derive their information from?
E.g. you could provide a Wikipedia post, which is referencing an article in the Times. Or perhaps the source contains a lot of synthetic data?
- While this seems really difficult to do technically, it would be a useful function, especially if the user might only have secondary sources. These questions seem to suggest a link to the service of credibility scoring.

Table 5. EAB recommendations for T6.1

EAB recommendation	Implementation measure/ Clarification
This service ‘aims to help journalists use GenAI tools responsibly while adhering to professional values and standards’. - Clarify your understanding of ‘responsibly’ - Clarify which professional values and standards are being considered	The concept of ‘responsibility’ is in discussion with the team already. It will be considered whether it could even be picked up in the simulation tool. Professional standards from journalistic ethics and frameworks on journalism and ethical AI use.
Create transparency on the AI models used to enable the simulator.	It is not clear how much ML will be used for this service, so far little to no use is foreseen. It is disclosed to the user, that some media of the simulator is generated by GenAI. If ML models are built, model cards will be created for these?
Clarify for the user how counter narratives are recommended and which data is used.	Explanation via button will be added
Users should be made aware that the sources they upload can be unreliable, so that they don’t blindly equate a ‘valid claim’ with a ‘fact’.	Explanation via button will be added
Create transparency on the AI models used to enable the counter narrative recommendation tool.	Explanation via button will be added

T6.2 Transparency Service for AI model cards

Comments/ Clarification

Varying levels of complexity are very much appreciated, because it seems to overcome the trade-off between transparency and usability.

“Ethical considerations” is too broad as a term and should be more narrowly focussed on the risks to beware of. Otherwise, the user might be overwhelmed by all the potential ethical benefits.

What are the AI elements exactly? → There are 3 models to ease the creation of model cards by creating an editable draft of the model card.

- One model to create an overview. There is an LLM-based model integrated that simplifies the model card description, based on a related research paper. This means that this is dependent on model documentation being available. This model simplifies the text from the documentation, but it does not yet take into consideration the specific audience.

- One model for retrieving specific information from the documentation e.g. use cases. This model is explicitly not creating new information, but retrieving information based on related research paper/ model documentation.
- One model that provides suggestions on the ethical risks: the developer provides a description of their model, and this model provides ideas on the ethical risks and proposes recommendations
 - o Current list of ethical risks: 1) misuse and dual-use concerns, 2) bias and fairness, 3) privacy and data security, 4) transparency and accountability, 5) societal impact, 6) environmental sustainability, 7) safety and reliability, 8) equity and accessibility, and 9) legal and regulatory challenges. --> *to be reviewed by CEPS*

Table 6. EAB recommendations for T6.2

EAB recommendation	Implementation measure/ Clarification
The level of detail to present in the model cards should be dependent on the preference of the user.	A consultation with media professionals (e.g. feedback survey) will be implemented to understand the user preference with regards to e.g. detail and technical complexity. Implementation of a function so that the user (developer) can instruct the AI model what the overview should look like considering user preferences.
The use of a chatbot to fill in model cards potentially invites biases, and thus we risk creating unreliable model cards. This should be monitored.	Implementing a user feedback mechanism to understand level of hallucinations and minimise this risk. Responsibility for the accuracy of the information lies with the creator of the model card. The created text is editable and should be reviewed by the creator of the model card who was made aware of this through the terms of service.
The use of a chatbot to fill in model cards can create a lack of transparency.	Model cards will be implemented for this AI-CODE service. The use of this AI service to create a model card should be made transparent to the end user (media professionals) in conjunction with the previous implementation measure so that trust in model cards created by this service is not reduced.
The field “ethical considerations” should be renamed “ethical risks”.	Wording has been changed.
The ethical risks should be presented so that they can be comprehended on one glance. For example, the cell specifies five ethical risks and each one can be opened in more detail with a drop-down list.	To be considered in conjunction with Row 1, but a drop down format should be implemented for most cells of the model cards, accommodating user preferences. The detailed explanations should be human-provided, the presentation of higher-level summaries can be considered to be supported by an AI model.
The potential biases in training and evaluation datasets should be made explicit in the cell of ethical risks and not just when the dataset is reported.	Biases in dataset are currently represented in the dataset overview, also in the model performance, and it is an ethical risk. In the representation layer for the user, the information can be linked in the ethical risks section (e.g. check the dataset description,

and model performance section), too.

T6.3 Personal Companion

Comments/ Clarification

The service has 3 integrated functionalities: Claim matching, fallacy detection, hate speech detection

- The fallacy detection functionality does not identify separate claims, but the logical fallacy directly, out of a list of 12 types of fallacies e.g. appeal to emotion. → It is LLM-based (fine-tuned Llama)
- Claim matching: The current state of the system is focused on identifying the main factual claim. However, the dataset used here is not final yet. The definition of 'claim' will be based on the data set and related research papers. Overall, the service identifies claims that may or may not be fact.
- Hate speech detection: Highlights that there is hate speech in the text out of 4 types of hate speech. It is mostly focussed on detecting strong language. → uses a smaller, fine-tuned model. ATC works on improving it.

The claim identification raised multiple issues, e.g. How do we define 'claim' in this service? Is this service distinguishing between claim and fact? How is the claim identifier dealing with weaker claims? E.g. statements "this could be/ something might be" What are the conditionals that the claim identifier can handle? Are we indicating how strong a claim is? How do we go about multiple claims wrapped in one sentence?

- The System is currently transitioning to a new data set. Answers to these questions depend on this new data set.

About the debunked database, the EAB asked if a user of a service could also see information on what was debunked about it and why, because this could enhance critical thinking and trust in the system.

- Yes. The database contains debunking reports, which include this information.

For which languages is this service available?

- English, so far.

Table 7. EAB recommendations for T6.3

EAB recommendation	Implementation measure/ Clarification
Re-word 'assist <i>young</i> professionals' to just 'professionals'. Neither age nor seniority level prevents one from committing logical fallacies.	Couldn't be changed for the past deliverable, but the wording has been adjusted. These things will be considered in communication about this service.
Promising that <i>all</i> logical fallacies will be detected is likely to create unrealistic expectations. The wording	Wording will be changed because this service captures 12 different types of logical fallacies.

should be more precise.	Transparency on these 12 different types can be offered through the UI. Currently, the user sees which type of fallacy is detected, not the full list. It is acknowledged that the service does not have a 100% accuracy rate. This will be reflected in the wording in description for deliverables and in user-facing communication.
Improve transparency on the reliability of this service.	Model cards will be implemented.
There should be more clarity for the user how the system operates and on its limitations, e.g. how is 'claim' defined, what type of logical fallacies are incorporated, and how conditionals are processed.	Model cards will be implemented. Additional textual information on the different functionalities can be incorporated in the UI, e.g. through a drop down menu or an info button.
It should be acknowledged that there is a difference between claim and fact.	The definition depends on the dataset. Alternative wording of "check-worthy claim" is explored.

Annex III – Ethical Design requirements

See also *D3.1 User-centred requirements definition*.

Table 8. List of ethical design requirements

#	Ethical requirements	Related to Trustworthy AI requirement
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system	Human agency and oversight (HAO)
HAO-02	The capabilities of the application shall be communicated clearly to the user.	HAO
HAO-03	The application shall include a detection and response mechanism for undesirable adverse effects of the AI system.	HAO
HAO-04	The application shall take into account the risk of stimulating dependency in users.	HAO
TSA-01	The application shall have a risk assessment system, i.e. risks, risk metrics, and risk levels of the AI system are defined.	Technical robustness & safety + Accountability (TSA)
TSA-02a	The accuracy, reliability, and reproducibility of the application shall be monitored, documented, and communicated to the user over time.	TSA
TSA-03	The data used to build the application shall be up-to-date, of high quality, complete and representative of the deployment environment.	TSA
TSA-04	The application shall be resilient to cyber-attacks.	TSA
PDG-01	The application shall minimize the amount of data collected on users.	Privacy and Data Governance (PDG)
TRA-01	The system input and output of the application shall be traceable through documentation.	Transparency (TRA)
TRA-02	The decisions of the application shall be understandable by the user (XAI) as far as reasonably achievable.	TRA
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI system.	TRA

FAI-01	The application shall have procedures in place that avoid unfair bias that apply to data, design, and output.	Fairness (FAI)
FAI-02	The application shall take the diversity and representativeness of the end user into account.	FAI
FAI-03	The application shall have a mechanism that allows for flagging issues related to bias, discrimination or poor performance.	FAI
ENV-01	The application shall identify, evaluate, and where possible mitigate environmental impacts (embed a carbon tracker) and impact on human working arrangements and democracy.	Environmental and societal wellbeing (ENV)
ACC-01	The application is designed so that it is auditable to third parties, i.e. the development process and sourcing of training data are traceable and the AI system's process, outcomes and impacts are logged consistently.	Accountability (ACC)
TSA-02b	There are processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated.	Consortium
ENV-02	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects.	Consortium
HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the use case.	Consortium

Annex IV – ALTAI Recommendations

Annex IV contains a table for each service with proposals on how to improve trustworthiness based on the ALTAI results from the M18 assessment. These measures draw on best practice examples from other services, or propose new approaches, if there are no suitable best practice examples, yet.

T4.1 Recommendations based on M18 ALTAI

Table 9. ALTAI Recommendations T4.1 (M18)

	Requirement for trustworthy AI	Best Practices from other services or Example Solutions
#1 Human Agency and oversight		
1	Implement procedures to avoid that end-users over-rely on the AI system	Show a disclaimer text when first logging into the service that clearly states limitations and intended use; provide model cards
2	Implement a ‘stop button’ to allow for a user’s request to be aborted.	Ensure that closing the browser window aborts the full process
#3 Privacy and Data Governance		
3	Establish mechanisms that allow for flagging issues related to privacy or data protection concerning the AI system	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text)
4	Implement GDPR thoroughly	Consult with the Data Protection Officer (DPO) at your organisation on whether a Data Protection Impact Assessment is necessary; privacy by design and default (e.g., encryption, pseudonymisation, aggregation, anonymisation); reduce the data you collect and store of users (data minimisation)
#4 Transparency		
5	Put in place measures to continuously* assess the quality of the input data to the AI system <i>*T4.1 isn’t using continuous learning, so input data doesn’t change the output quality. This requirement is interpreted broadly</i>	Evaluate the accuracy of moderation decisions that underpin the crowdsourced information you provide through your service, e.g. through random sampling

6	Explain the decision of the AI system to the users	Provide users with the option to learn more about underlying systems through an 'info'-button; make detailed model card available through a tab or link in the UI
7	Survey the users to ask them whether they understand the decision(s) of the AI system	In the upcoming co-creation sessions and user-based evaluation in WP8 include questions to get a sense of their understanding; survey early adopters
8	Establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system	Show a disclaimer text in the beginning which includes this information; make detailed model card available
9	Communicate the technical limitations and potential risks of the AI system to end-users, such as its level of accuracy and/or error rates	Show a disclaimer text in the beginning which includes this information; provide an info-button for the different models; make detailed model card available
#5 Diversity, non-discrimination and fairness		
10	Test for specific target groups or problematic use cases	Test a range of different posts and requests that could be entered into the system
11	Assess and put in place processes to test and monitor for potential biases of the AI system	Test for biases in datasets and model performance; allow for user flagging of these issues via mechanism (such as email, see #1)
12	Consider diversity in end-users	Review the UI for its accessibility; consider accessibility-by-design and universal design principles
13	Ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system	Provide an up-to-date email address as well as instructions to the user that such issues can be flagged through this mechanism
14	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users	Create a stakeholder mapping (e.g. based on a collective team brainstorming); consider how these subjects could potentially be impacted; implement a risk management system
15	Consider thorough implementation of Universal Design and test the accessibility of your service.	Find additional resources that help you to improve accessibility, such as the Web Content Accessibility Guidelines (WCAG)
#6 Societal and Environmental Well-being		
16	Evaluate the environmental impact of the AI system's development, deployment and/or use	Implement a carbon tracker; monitor energy use

17	Ensure that the impacts of the AI system on working arrangements are well understood	Ensure that the work impact is understood through co-creation sessions and user-based validation; talk to early adopters to understand how the tool is impacting their workflow
#7 Accountability		
18	Improve auditability	Log the operations during deployment
19	Implement a redress by design mechanism	Ensure that outputs of the AI system can be flagged through a mechanism and subsequently investigated and improved by you

T4.2 Recommendations based on M18 ALTAI

Table 10. ALTAI Recommendations T4.2 (M18)

	Requirement for trustworthy AI	Best Practices from other services or Example Solutions
#1 Human Agency and oversight		
1	Put in place any procedure to avoid that the system inadvertently affects human autonomy	Strengthen communication on limitations of the system against over-reliance; understand through user-based validation the concrete risk here, e.g. does the service significantly bias the assessment of media professionals?
2	Give specific training to those exercising oversight over the system	Ensure training materials are prepared for future deployers and users
#3 Privacy and Data Governance		
3	Establish mechanisms that allow for flagging issues related to privacy or data protection concerning the AI system	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text)
4	Implement GDPR thoroughly	Consult with the Data Protection Officer (DPO) – or ethical committee – at your organisation on whether a Data Protection Impact Assessment is necessary
#4 Transparency		
5	Communicate the technical limitations and potential risks of the AI system to end-users, e.g. accuracy and error rates	Show a disclaimer text in the beginning which includes this information; provide an info-button for the different models; make detailed model card available
#5 Diversity, non-discrimination and fairness		

6	Establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system that considers diversity and representativeness of subjects in the data	For the bot-detector, consider performance indicators across different types of social media users, e.g. non-native speakers, subcultures, etc.
7	Improve your understanding of the data, model and performance with regards to bias	Research and use publicly available technical tools that are state-of-the-art
8	Assess and put in place processes to test and monitor for potential biases of the AI system	Test for biases in datasets and model performance; allow for user flagging of these issues via mechanism (such as email, see #1)
9	Consider diversity in end-users and implement and test accessibility of your service.	Review the service for its accessibility; consider accessibility-by-design and universal design principles; find additional resources that help you to improve accessibility, such as the Web Content Accessibility Guidelines (WCAG)
10	Ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system	Provide an up-to-date email address as well as instructions to the user that such issues can be flagged through this mechanism
11	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users	Create a stakeholder mapping (e.g. based on a collective team brainstorming); consider how these subjects could potentially be impacted; implement a risk management system
#6 Societal and Environmental Well-being		
12	Evaluate the environmental impact of the AI system's development, deployment and/or use	Implement a carbon tracker; monitor energy use
13	Reduce the environmental impact	Select small, energy-efficient models where possible, encourage organisational initiatives
#7 Accountability		
14	Improve auditability	Log the operations during deployment
15	Implement a redress by design mechanism	Ensure that outputs of the AI system can be flagged through a mechanism and subsequently investigated and improved by you

T4.3 Recommendations based on M18 ALTAI

Table 11. ALTAI Recommendations T4.3 (M18)

	Requirement for trustworthy AI	Best Practices from other services or Example Solutions
#1 Human Agency and oversight		
1	Put in place any procedure to avoid that the system inadvertently affects human autonomy	Strengthen communication on limitations of the system against over-reliance; understand through user-based validation the concrete risk here, e.g. does the service significantly bias the assessment of media professionals?
2	Establish detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text); review output accuracy of system periodically
3	Give specific training to those exercising oversight over the system	Ensure training materials are prepared for early adopters and future deployers/users
#3 Privacy and Data Governance		
4	Establish mechanisms that allow for flagging issues related to privacy or data protection concerning the AI system	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text)
5	Implement GDPR thoroughly	Consult with the Data Protection Officer (DPO) at your organisation on whether a Data Protection Impact Assessment is necessary
#4 Transparency		
6	Put in place measures to continuously assess the quality of the input data to the AI system	Show a disclaimer text in the beginning which includes information on how quality of input impacts on quality of output; periodically review performance of models with different quality input
#5 Diversity, non-discrimination and fairness		
7	Improve your understanding of the data, model and performance with regards to bias	Research and use publicly available technical tools, that are state-of-the-art
8	Assess and put in place processes to test and monitor for potential biases of the AI system	Test for biases in datasets and model performance; allow for user flagging of these issues via mechanism (such as email, see #1)
9	Consider diversity in end-users and implement and test accessibility of your service.	Review the service for its accessibility; consider accessibility-by-design and universal design principles; find additional resources that help you to improve accessibility, such as the Web Content Accessibility Guidelines (WCAG)

10	Ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system	Provide an up-to-date email address as well as instructions to the user that such issues can be flagged through this mechanism
11	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users	Create a stakeholder mapping (e.g. based on a collective team brainstorming); consider how these subjects could potentially be impacted; implement a risk management system
#6 Societal and Environmental Well-being		
12	Evaluate the environmental impact of the AI system's development, deployment and/or use	Implement a carbon tracker; monitor energy use
13	Reduce the environmental impact	Choose small, energy-efficient models where possible; encourage organisational initiatives for energy efficiency
#7 Accountability		
14	Improve auditability	Reconstruct how AI models have been trained to improve reproducibility of results
15	Implement a redress by design mechanism	Ensure that outputs of the AI system can be flagged through a mechanism and subsequently investigated and improved by you

T6.1A Recommendations based on M18 ALTAI

Table 12. ALTAI Recommendations T6.1A (M18)

	Requirement for trustworthy AI	Best Practices from other services or Example Solutions
#1 Human Agency and oversight		
#3 Privacy and Data Governance		
1	Establish mechanisms that allow for flagging issues related to privacy or data protection concerning the AI system	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text)
#4 Transparency		
#5 Diversity, non-discrimination and fairness		
2	Test for specific target groups or problematic use cases	Test a range of different posts and requests that could be entered into the system

3	Assess and put in place processes to test and monitor for potential biases of the AI system	Test for biases in datasets and model performance; allow for user flagging of these issues via mechanism (such as email, see #1)
4	Consider diversity in end-users	Review the UI for its accessibility; consider accessibility-by-design and universal design principles
5	Ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system	Provide an up-to-date email address as well as instructions to the user that such issues can be flagged through this mechanism
6	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users	Create a stakeholder mapping (e.g. based on a collective team brainstorming); consider how these subjects could potentially be impacted; implement a risk management system
7	Consider thorough implementation of Universal Design and test the accessibility of your service.	Find additional resources that help you to improve accessibility, such as the Web Content Accessibility Guidelines (WCAG)
#6 Societal and Environmental Well-being		
#7 Accountability		
8	Improve auditability	Log the operations during deployment or encourage users to save screenshots when logging problematic interactions

T6.1B Recommendations based on M18 ALTAI

Table 13. ALTAI Recommendations T6.1B (M18)

	Requirement for trustworthy AI	Best Practices from other services or Example Solutions
#1 Human Agency and oversight		
#3 Privacy and Data Governance		
1	Establish mechanisms that allow for flagging issues related to privacy or data protection concerning the AI system	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text)
#4 Transparency		
#5 Diversity, non-discrimination and fairness		
2	Improve your understanding of the data, model and performance with regards to bias	Research and use publicly available technical tools, that are state-of-the-art

3	Assess and put in place processes to test and monitor for potential biases of the AI system	Test for biases in datasets and model performance; allow for user flagging of these issues via mechanism (such as email, see #1)
4	Consider diversity in end-users and implement and test accessibility of your service.	Review the service for its accessibility; consider accessibility-by-design and universal design principles; find additional resources that help you to improve accessibility, such as the Web Content Accessibility Guidelines (WCAG)
5	Ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system	Provide an up-to-date email address as well as instructions to the user that such issues can be flagged through this mechanism
6	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users	Create a stakeholder mapping (e.g. based on a collective team brainstorming); consider how these subjects could potentially be impacted; implement a risk management system
#6 Societal and Environmental Well-being		
7	Evaluate the environmental impact of the AI system's development, deployment and/or use	Implement a carbon tracker; monitor energy use
#7 Accountability		
8	Implement a system for identifying and dealing with conflict of values	Assess whether there are any conflicts of interest among the developing team

T6.2 Recommendations based on M18 ALTAI

Table 14. ALTAI Recommendations T6.2 (M18)

	Requirement for trustworthy AI	Best Practices from other services or Example Solutions
#1 Human Agency and oversight		
1	Implement a 'stop button' to allow for a users' request to be aborted.	Ensure that closing the browser window aborts the full process
#3 Privacy and Data Governance		
2	Establish mechanisms that allow for flagging issues related to privacy or data protection concerning the AI system	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text)
3	Implement GDPR thoroughly	Consult with the Data Protection Officer (DPO) at your organisation on whether a Data Protection Impact Assessment is necessary; privacy by design and default

		(e.g., encryption, pseudonymisation, aggregation, anonymisation); reduce the data you collect and store of users (data minimisation)
#4 Transparency		
4	Put in place measures to continuously assess the quality of the input data to the AI system	Ensure that users are made aware that the quality of material provided for the service for generating model cards impacts quality of model cards generated
5	Explain the decision of the AI system to the users	Provide users with the option to learn more about underlying systems through an 'info'-button; make detailed model card available through a tab or link in the UI
6	Survey the users to ask them whether they understand the decision(s) of the AI system	In the upcoming co-creation sessions and user-based evaluation in WP8 include questions to get a sense of their understanding; survey early adopters
7	Establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system	Show a disclaimer text in the beginning which includes this information; make detailed model card available
8	Communicate the technical limitations and potential risks of the AI system to end-users, such as its level of accuracy and/or error rates	Show a disclaimer text in the beginning which includes this information; provide an info-button for the different models; make detailed model card available
#5 Diversity, non-discrimination and fairness		
9	Test for specific target groups or problematic use cases	Test a range of different posts and requests that could be entered into the system
10	Assess and put in place processes to test and monitor for potential biases of the AI system	Test for biases in datasets and model performance; allow for user flagging of these issues via mechanism (such as email, see #1)
11	Consider diversity in end-users	Review the UI for its accessibility; consider accessibility-by-design and universal design principles
12	Ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system	Provide an up-to-date email address as well as instructions to the user that such issues can be flagged through this mechanism
13	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users	Create a stakeholder mapping (e.g. based on a collective team brainstorming); consider how these subjects could potentially be impacted; implement a risk management system
14	Consider thorough implementation of Universal Design and test the accessibility of your service.	Find additional resources that help you to improve accessibility, such as the Web Content Accessibility Guidelines (WCAG)

#6 Societal and Environmental Well-being		
15	Evaluate the environmental impact of the AI system's development, deployment and/or use	Implement a carbon tracker; monitor energy use
16	Ensure that the impacts of the AI system on working arrangements are well understood	Ensure that the work impact is understood through co-creation sessions and user-based validation; talk to early adopters to understand how the tool is impacting their workflow
#7 Accountability		
17	Implement a redress by design mechanism	Ensure that outputs of the AI system can be flagged through a mechanism and subsequently investigated and improved by you

T6.3 Recommendations based on M18 ALTAI

Table 15. ALTAI Recommendations T6.3 (M18)

	Requirement for trustworthy AI	Best Practices from other services or Example Solutions
#1 Human Agency and oversight		
#3 Privacy and Data Governance		
1	Establish mechanisms that allow for flagging issues related to privacy or data protection concerning the AI system	Provide an up-to-date email address as well as instructions to the users that such issues can be flagged through this mechanism (e.g. in disclaimer text)
#4 Transparency		
2	Put in place measures to continuously* assess the quality of the input data to the AI system <i>*T6.3 isn't using continuous learning, so input data doesn't change the output quality. This requirement is interpreted broadly</i>	Ensure that users are aware that output quality depends (in part) on quality of text provided for analysis and the limitations of this tool on poor quality text.
3	Communicate the technical limitations and potential risks of the AI system to end-users, such as its level of accuracy and/or error rates	Include this information in the disclaimer text; provide an info-button for the different models; make detailed model card available
#5 Diversity, non-discrimination and fairness		

4	Test for specific target groups or problematic use cases	Test a range of different posts and requests that could be entered into the system
5	Assess and put in place processes to test and monitor for potential biases of the AI system	Test for biases in datasets and model performance; allow for user flagging of these issues via mechanism (such as email, see #1)
6	Consider diversity in end-users	Review the UI for its accessibility; consider accessibility-by-design and universal design principles
7	Ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system	Provide an up-to-date email address as well as instructions to the user that such issues can be flagged through this mechanism
8	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the end-users	Create a stakeholder mapping (e.g. based on a collective team brainstorming); consider how these subjects could potentially be impacted; implement a risk management system
9	Consider thorough implementation of Universal Design and test the accessibility of your service.	Find additional resources that help you to improve accessibility, such as the Web Content Accessibility Guidelines (WCAG)
#6 Societal and Environmental Well-being		
10	Evaluate the environmental impact of the AI system's development, deployment and/or use	Implement a carbon tracker; monitor energy use
11	Provide training materials for re- and up-skilling measures	In collaboration with early adopters, identify the need for training materials
#7 Accountability – No recommendations		