TRUSTWORTHY AI ASSESSMENT LIST[1]
ARCHIBOT 3.0

Table of Contents

[1] Check list based on "THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI) for self-assessment, INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION, 17 July 2020, Book: ISBN 978-92-76-20009-3

Name of the solution: Archibot 3.0

Deployed package: ep-submit-nlp-search-v8.1

Purpose of the information system: Providing an easy access through natural language to public historical archived documents from the European Parliament in a multilingual context. The purpose of this solution is to enable various strategies for searching information in a multilingual context, making use of a retrieval augmented generation solution.

Type of AI solution: Retrieval Augmented Generation with generative AI

Access to the system: https://archidash.europarl.europa.eu/ep-archives-anonymous-dashboard , select in "Dashboard" the option "content-analysis" and then "Ask the EP archives", access to the description of the system use: https://historicalarchives.europarl.europa.eu/en/sites/historicalarchive/home/cultural-heritage-collections/news/ai-dashboard.html

Respect of "Guidelines, Use of publicly available Artificial Intelligence tools for Parliament staff", note CSG D(2024)14346 of 30 April 2024

Principle n°1: Non-Disclosure and personal data protection
Comply with the principle: **Yes** - No

Principle n°2: Content responsibility
Comply with the principle: **Yes** - No

Principle n°3: Transparency and Compliance
Comply with the principle: **Yes** - No

Principle n°4: Autonomy and Business Continuity
Comply with the principle: **Yes** - No

REQUIREMENT #1 Human Agency and Oversight

Purpose: AI systems should support human agency and human decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both: act as enablers for a democratic, flourishing and equitable society by supporting the user's agency; and uphold fundamental rights, which should be underpinned by human oversight. In this section, AI systems are assessed in terms of their respect for human agency and autonomy as well as human oversight.

Human Agency and Autonomy

This subsection deals with the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision-making processes, for example, algorithmic decision support systems, risk analysis/prediction systems (recommender systems, predictive policing, financial risk analysis, etc.). It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans. Finally, it deals with the effect of AI systems on human affection, trust and (in)dependence.

• Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society? *Yes, the system is designed to interact with humans*.

• Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision? *A disclaimer states clearly that the answer provided to the question results form an algorithm and is not authoritative (disclaimer section).*

• Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision? *A specific document is provided to explain how to use the solution and how the output is build (how to use section).*

• Are end-users or subjects informed that they are interacting with an AI system? *This is very explicitly mentioned in the how to use section.*

• Could the AI system affect human autonomy by generating over-reliance by end-users? *The section related to how to use the system provides an explicit declaration on the way the question is handled, the selection of documents is performed and the answer is generated. The user is invited to read all primary sources that are provided to establish its own interpretation of these ones. A section on the use of the system is provided to mitigate this risk. All documents that are used to generate the answer are available to the user in the tab "Advanced content search"*

• Could the AI system affect human autonomy by interfering with the end user's decision-making process in any other unintended and undesirable way? *No, it is stated that the system provides an answer considering a calculated set of documents pertinent for the answer.*

• Did you put in place any procedure to avoid that the AI system inadvertently affects human autonomy? *No specific procedure is set while only managing a non-authoritative answer to a question provided by a user. Moreover, as stated in previous answers, it is made clear the answer is provided by an AI solution.*

• Does the AI system simulate social interaction with or between end-users or subjects? *No, it is stated in the disclaimer that this system is not a chatbot and that the questions are not stored.*

• Depending on which risks are possible or likely, please answer the questions below:
    o Did you take measures to deal with possible negative consequences for end-users or subjects in case they develop a disproportionate attachment to the AI System? *This system is only providing access to legislative document. It is intended to be used by young researchers or researchers, eventually by citizens interested in the European integration. While no interactions are maintained but the system is only providing an answer to a question, disproportionate attachment is not expected.*
    o Did you take measures to minimise the risk of addiction? *No interactions are recorded, no invitation for exchanges are proposed.*
    o Did you take measures to mitigate the risk of manipulation? *The system is providing an answer to a question from the user. The answer is established based on official documents from the European parliament. The primary sources used to provide an answer are systematically provided with the answer.*

Human Oversight

This subsection helps to self-assess necessary oversight measures through governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approaches. Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system. Human-on-the-loop refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. Human-in-command refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the AI system in any particular situation. The latter can include the decision not to use an AI system in a particular situation to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by an AI system.

• Please determine whether the AI system (choose as many as appropriate):
    o Is a self-learning or autonomous system: *No*
    o Is overseen by a Human-in-the-Loop: *Yes, each document have been qualified by a human and the LLM is used in RAG mode. It is foreseen to implement in a short time frame Model evalutation to automate the evaluation of the answer provided to the test prompt set. Currently the prompt set is applied manually to verify the adequacy of the answer to a set of prompts.*
    o Is overseen by a Human-on-the-Loop: *Yes, no automatic code generation have been used and steps performed by the solution are explicitly provided to the end user.*
    o Is overseen by a Human-in-Command: *Yes, the solution benefits from the AI constitutional approach of the LLM and safe guard rules implemented in the LLM to protect against undesired topics.*

• Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight? *Yes, the staff in charge of the design and the monitoring of the system have received a training on the use and implementation of AI systems. They have been provided with this assessment to reinforce the importance and the conditions of an ethical approach.*

• Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject? *The Large Language Model (LLM) implemented is based on an AI Constitutional approach*[2]*. This ensures that answers provided respects the values provided by the Declaration of Human Rights from the United Nations. The most undesirable effect estimated is a surprised on the technical content of an answer, reason why all the primary resources used to establish the answers are provided. Safe guard rules are implemented in the LLM to prevent the use of topics which are not compliant with the AI constitutional approach.*

• Did you ensure a 'stop button' or procedure to safely abort an operation when needed*? The system can be stopped at any time from the monitoring console accessed by the administrator of the system.*

• Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system? *The system is not designed to learn from questions or answers. When Claude is used on AWS Bedrock the data is not transmitted to the model provider (Anthropic) and hence no self-learning is possible: https://aws.amazon.com/bedrock/security-compliance/*

[2] https://www.anthropic.com/news/claudes-constitution

REQUIREMENT #2 Technical Robustness and Safety

Purpose: A crucial requirement for achieving Trustworthy AI systems is their dependability (the ability to deliver services that can justifiably be trusted) and resilience (robustness when facing changes). Technical robustness requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimising unintentional and unexpected harm as well as preventing it where possible. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner. The questions in this section address four main issues: 1) security; 2) safety; 3) accuracy; and 4) reliability, fall-back plans and reproducibility.

Resilience to Attack and Security

• Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use? [redacted] *Jailbreaking and prompt injections occur when users craft specific prompts that exploit vulnerabilities in the model's training, aiming to generate inappropriate or harmful content.* [redacted] *Learn more here; https://docs.anthropic.com/en/docs/mitigating-jailbreaks-prompt-injection*

• Is the AI system certified for cybersecurity (e.g. the certification scheme created by the Cybersecurity Act in Europe) or is it compliant with specific security standards? [redacted]

• Did you assess potential forms of attacks to which the AI system could be vulnerable? [redacted]

• Did you consider different types of vulnerabilities and potential entry points for attacks such as:

- Data poisoning (i.e. manipulation of training data): [redacted]
- Model evasion (i.e. classifying the data according to the attacker's will): *all documents used to generate the answer are public archived documents.*
- Model inversion (i.e. infer the model parameters): [redacted]

• Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle? *The system is hosted in cloud premises. It takes benefits from the capacity of the provider. Integrity of documents are ensured by* [redacted] *The robustness is provided by* [redacted]

[redacted]

• Did you red-team/pentest the system? [redacted]

• What length is the expected timeframe within which you provide security updates for the AI system? [redacted]

General Safety

• Did you put in place a process to continuously measure and assess risks? [redacted]

• Did you inform end-users and subjects of existing or potential risks? *Yes, it is mentioned that all answers provided are not authoritative and they are invited to consult all primary sources used to generate the answer.*

• Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system? [redacted]

• Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system? [redacted]

• Did you align the reliability/testing requirements to the appropriate levels of stability and reliability? [redacted]

• Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')? [redacted]

• Did you develop a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness and safety? [redacted]

Accuracy

• Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences? *The probability of such a risk is extremely low and with no candidate situation identified.*

• Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in? *The training data set of the LLM is compliant with a*

Constitutional AI and available on the site of the provider (Anthropic Claude 3.5 Sonnet). The training data is not publicly available, but learn more about the types of data used here: https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training. This version is embedded when made available to improve the accuracy of the understanding of a demand. The document set use to generate the answer is an official set of digital copies from the European parliament from the archive management system.

• Did you put in place a series of steps to monitor, and document the AI system's accuracy? *A test set of questions has been defined to evaluate the answer and the version system was challenged with ChaGPT 3.5 and Titan Text in terms of answer accuracy. It provided better results with Claude 3.0. The translation from the prompt and the answer to the user are provided through Anthropic Claude 3.5 Sonnet, resulting from the deployment of the last version of this LLM. All data used to described the documents in the archive management system for which these documents are extracted are in French.*

• Did you consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects? *The design of the system is made to avoid such a bias, except if this behaviour would be intentionally included within the LLM. The test set applied to qualify the system was establish to evaluate this type of operation, which has not be proven or detected.*

• Did you put 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? *A disclaimer and how to use sections are provided in this purpose.*

Reliability, Fall-back plans and Reproducibility

• Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals? *The question test set is applied before each new version is deployed to evaluate the impact of the new version on the answers provided.*

• Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility? *Reproducibility is tested with the test question set (30 questions).*

• Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system? *The question test set is applied systematically.*

• Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them? *If an unexpected behaviour is identified, it is expected to reproduce the same case (the same question) having activated the traces at the level of the console to identify potential issues and to correct them.*

• Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score? *The answers from the system on the question test set were qualified by experts as acceptable answers. These ones may not be complete, however this is part of the design of the system (only the 10 documents matching the best to the system are used as a context to generate the answer) and communicated clearly to the user. This point is revised too with the application of the test prompt set that will be automated, making use of Model evaluation.*

• Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function? *The system is not using continual learning. Only new documents to be published according the legal requirements (Council Regulation 1983/354) are provided and major versions of the LLM as mentioned above. In addition, No data transmitted to the model on AWS Bedrock is transmitted to Anthropic. AWS Bedrock does not use customer prompts and completions to improve the model.*

REQUIREMENT #3 Privacy and Data Governance

Purpose: Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

Privacy

This subsection helps to self-assess the impact of the AI system's impact on privacy and data protection, which are fundamental rights that are closely related to each other and to the fundamental right to the integrity of the person, which covers the respect for a person's mental and physical integrity.

• Did you consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection? *Yes, the system is only using European Parliament public archived data, which are by definition public.*

• Depending on the use case, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system? *All data are public.*

Data Governance

This subsection helps to self-assess the adherence of the AI system('s use) to various elements concerning data protection.

• Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)? *The documents used to generate the answers are public archived documents and the LLM is trained according to an AI Constitutional approach to prevent this type of processing. The Claude models are specifically trained to respect privacy: one of the constitutional "principles" at the heart of Claude, based on the Universal Declaration of Human Rights, is to choose the response that is most respectful of everyone's privacy, independence, reputation, family, property rights, and rights of association.*

• Did you put in place any of the following measures some of which are mandatory under the Regulation (EU) 2018/1725 on personal data protection, or a non-European equivalent?

- Data Protection Impact Assessment (DPIA): *Yes, this system is related to the declaration n2 of the European Parliament register, all data are public archived data;*
- Designate a Data Protection Officer (DPO) and include them at an early state in the development, procurement or use phase of the AI system: *Yes, the DPO was informed about this system and functionalities to ensure its alignment with the data processing;*
- Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications): [redacted]

- Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation): *The documents used to generate the answers are public archived documents;*
- Data minimisation, in particular personal data (including special categories of data): *The documents used to generate the answers are public archived documents.*

• Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data? *The training set as been evaluated according to its compliance to the AI Constitutional approach to mitigate this risk.*

• Did you align the AI system with relevant standards (e.g. ISO25, IEEE26) or widely adopted protocols for (daily) data management and governance? The system aligns with the accessibility of public archives (Council Regulation 1983/354).

REQUIREMENT #4 Transparency

Purpose: A crucial component of achieving Trustworthy AI is transparency, which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system.

Traceability

This subsection helps to self-assess whether the processes of the development of the AI system, i.e. the data and processes that yield the AI system's decisions, is properly documented to allow for traceability, increase transparency and, ultimately, build trust in AI in society.

• Did you put in place measures to continuously assess the quality of the input data to the AI system? *all documents provided to the system have been verified by a human through a quality control of all data exposed.*

• Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)? *Yes, all documents used to generate an answer are provided below the answer and accessible (even by download) to the user.*

• Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system? *One system is used to search for the maximum 10 pertinent documents for a question (distance calculation between the tokens of the question and the indexes of the overall document data set, OpenSearch). One system is used elaborate the answer (LLM) based on the documents identified (Anthropic/Claude 3.5 Sonnet).*

• Did you put in place measures to continuously assess the quality of the output(s) of the AI system? *The test question set is applied before each releases. This will be automated by the use of Model evaluation on the test prompt data set.*

• Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system? *The traces can be found in the log files in case of need. Logging data are available in CloudTrail.*

Explainability

This subsection helps to self-assess the explainability of the AI system. The questions refer to the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes. Explainability is crucial for building and maintaining users' trust in AI systems. AI driven decisions – to the extent possible – must be explained to and understood by those directly and indirectly affected, in order to allow for contesting of such decisions. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black boxes' and require special attention. In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities) may be required, provided that the AI system as a whole respects fundamental rights. The degree to which explainability is needed depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life.

• Did you explain the decision(s) of the AI system to the users? *This is clearly stated in the disclaimer and how to use sections.*

• Do you continuously survey the users if they understand the decision(s) of the AI system? *A group of external test users (University of Wien, Austria) is maintained to improve the quality of the system is weaknesses are identified.*

Communication

This subsection helps to self-assess whether the AI system's capabilities and limitations have been communicated to the users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy as well as its limitations.

• In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human? *This is clearly stated in the disclaimer.*

• Did you communicate the benefits of the AI system to users? *Yes, in the how to use section.*

• Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates? *Yes, in the disclaimer section.*

• Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system? *Yes, just close to the prompt zone.*

REQUIREMENT #5 Diversity, Non-discrimination and Fairness

Purpose: In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a non transparent market. Identifiable and discriminatory bias should be removed in the collection phase where possible. AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance.

Avoidance of Unfair Bias

• Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design? *The input data are all official public archived documents from the European Parliament. The documents are based on legislative terms (all documents from a specific term are provided). These documents are all the ones in possession of the European Parliament for the corresponding type of document and legislature. Type of documents are motion for resolutions, parliamentary questions and answers, adopted texts. No selection is made on any specific criterion. The algorithm to select the documents is a distance calculation between the tokens of the question and the indexes (tokens) resulting from the indexing of the documents (tokens extracted from the documents)*[3].

• Did you test for specific target groups or problematic use cases? *Documents to be searched in are all the ones in possession of the archives of the European Parliament for the type of comments and legislature proposed.*

• Did you research and use publicly available technical tools, that are state-ofthe-art, to improve your understanding of the data, model and performance? *The document set is a fixed set of documents without evolution. All of them have controlled by a human*[4].

• Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)? *The document set used to provide the answers is a complete set of type of documents for a legislature; no selection is made in the document set to be searched in.*

• Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data? *End-users are mainly young researches, researchers and archivists.*

[3] Additional detail on Anthropic's approach to mitigating bias in the models: https://www.anthropic.com/news/evaluating-and-mitigating-discrimination-in-language-model-decisions

[4] Anthropic has documentation here to support in how to use Claude to get highest quality outputs: https://docs.anthropic.com/en/docs/prompt-engineering

• Did you put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system? *AI designers have been trained on ethical issues for AI systems.*

• Did you establish clear steps and ways of communicating on how and to whom such issues can be raised? *Such a situation has not been identified while documents proposed are legislative documents, resulting from the official positions of Members of the European Parliament (motion for resolutions, parliamentary question, adopted texts) or from a European institution (answers to parliamentary questions). These documents proposed have been subject to a verification for acceptance according to the Rules of the European Parliament once tabled. Then they have been the matter of an evaluation by the originating organ with the European Parliament following the retention schedule and the decision applicable following this schedule (destroy or preserve).*

• Did you identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects? *No subjects have been identified, according to the previous answer.*

• Did you consider other definitions of fairness before choosing this one? *No*

• Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities? *Young researchers, researchers, archivists and librarians in European Parliaments (ECPRD group and interinstitutional archive group) and member of the Centre for Innovation in Parliaments (Interparliamentary Union) were informed about this system.*

• Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness? *This assessment is used to evaluate the implementation of fairness and the answer provided by the system to the questions.*

• Did you establish mechanisms to ensure fairness in your AI system? *Expert review of the questions for the question test set.*

Accessibility and Universal Design

Particularly in business-to-consumer domains, AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards. This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies.

• Did you ensure that the AI system corresponds to the variety of preferences and abilities in society? *This system is designed for a business-to-business context more than a business-to-consumer. Users expected are young researchers, researchers, archivists or historians.*

• Did you ensure that information about, and the AI system's user interface of, the AI system is accessible and usable also to users of assistive technologies (such as screen readers)? *This is under consolidation.*

• Did you involve or consult with end-users or subjects in need for assistive technology during the planning and development phase of the AI system? *Yes, the site including this system is monthly crawled by SiteImprove*.

• Did you ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable? *As much as possible in terms of clarity and simplicity.*

• Did you assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects? *Engagement is ensured with the test user set*.

• Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI system? *As mentioned in the Avoidance of unfair bias, the document set results from all official positions of Members during their legislative activities.*

• Did you assess the risk of the possible unfairness of the system onto the end-user's or subject's communities? *Yes, as stated in the previous sections of this document.*

Stakeholder Participation

In order to develop Trustworthy AI, it is advisable to consult stakeholders who may directly or indirectly be affected by the AI system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.

• Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development? *Young researchers and researchers from the Universities of Luxembourg and Wien, archivists and librarians in European Parliaments (ECPRD group and interinstitutional archive group) and member of the Centre for Innovation in Parliaments (Interparliamentary Union) were informed about this system and asked for their feed-back.*

REQUIREMENT #6 Societal and Environmental Well-being

Purpose: In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle. Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or negatively impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could equally affect peoples' physical and mental well-being. The effects of AI systems must therefore be carefully monitored and considered. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, for instance the Sustainable Development Goals. Overall, AI should be used to benefit all human beings, including future generations. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation or democratic voting systems or pose a systemic threat to society at large.

Environmental Well-being

This subsection helps to self-assess the (potential) positive and negative impacts of the AI system on the environment. AI systems, even if they promise to help tackle some of the most pressing societal concerns, e.g. climate change, must work in the most environmentally friendly way possible. The AI system’s development, deployment and use process, as well as its entire supply chain, should be assessed in this regard (e.g. via a critical examination of the resource usage and energy consumption during training, opting for less net negative choices). Measures to secure the environmental friendliness of an AI system’s entire supply chain should be encouraged.

• Which potential impact(s) do you identify? *The only impact identified is the energy consumption for IT equipment to produce the answer. The LLM is provided through a stable version (no permanent learning) in the cloud environment and documents are indexed only when a type of document is available for a global legislature to avoid multiple trainings. The cloud provider mentions (available on AWS site) that the energy used is 90% produced by renewal energy sources. Other sustainable initatives are available on their site.*

• Did you define measures to reduce the environmental impact of the AI system throughout its lifecycle? *Minimising the training due to the availability of new documents. Minimizing the deployment of LLMs to major versions.*

Impact on Work and Skills

AI systems may fundamentally alter the work sphere. They should support humans in the working environment, and aim for the creation of meaningful work. This subsection helps self-assess the impact of the AI system and its use in a working environment on workers, the relationship between workers and employers, and on skills.

• Does the AI system impact human work and work arrangements? *The aim of the system is to augment a human activity (people having the European integration as a subject matter).*

• Did you pave the way for the introduction of the AI system in your organisation by informing and consulting with impacted workers and their representatives (trade unions, (European) work

councils) in advance? *Yes, this project is part of a project in the Strategic Execution Framework 2019-2024 from the European Parliament.*

• Did you ensure that workers understand how the AI system operates, which capabilities it has and which it does not have? *Yes through an explicit disclaimer and how to use section.*

• Did you take measures to counteract de-skilling risks? *The system proposes an answer to a question from the user to highlight information contained in public archived documents. The objective is to raise awareness. This does not replace eventually a consultation by the user of all public archived documents proposed. This awareness is raised outside the European Union for these documents an access method as promoting a democratic opening in a legislative activity.*

• Did you provide training opportunities and materials for re- and up-skilling? *A section is proposed on how to submit a question.*

Impact on Society at large or Democracy

This subsection helps to self-assess the impact of an AI system from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration, particularly in situations relating to the democratic processes, including not only political decision-making but also electoral contexts (e.g. when AI systems amplify fake news, segregate the electorate, facilitate totalitarian behaviour, etc.).

• Did you assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large? *The purpose is to raise awareness into topics that are unknown or difficult to find. This reinforces to notion of accountability of the European Parliament.*

• Did you take action to minimize potential societal harm of the AI system? *The LLM used is designed to respect human right values (see the points above on AI constitutional approach on Claude models).*

• Did you take measures that ensure that the AI system does not negatively impact democracy? *The system is supporting the transparency and accountability of the European Parliament according to the Council Regulation 1983/354.*

REQUIREMENT #7 Accountability

Purpose: The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems. This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties. When unjust or adverse impacts occur, accessible mechanisms for accountability should be in place that ensure an adequate possibility of redress.

Auditability

This subsection helps to self-assess the existing or necessary level that would be required for an evaluation of the AI system by internal and external auditors. The possibility to conduct evaluations as well as to access records on said evaluations can contribute to Trustworthy AI. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available.

• Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)? *The source code has been opened to the cybersecurity directorate. The LLM data training set is established by its provider, Anthropic / Claude 3.5 Sonnet. The most important criterion in this case is the AI Constitutional approach used. Logging data are available in CloudTrail.*

• Did you ensure that the AI system can be audited by independent third parties*? The source code results mainly from service integration on the platform of the cloud service provider. No external providers were involved in the design and implementation of the system, only official staff. Access to the source code could be provided according to the regulation in force for this topic.*

Risk Management

Both the ability to report on actions or decisions that contribute to the AI system's outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system.

When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to safety and ethical principles, including fundamental rights. Any decision about which trade-off to make should be well reasoned and properly documented. When adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.

• Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures? *The AI Competence Centre, as an independent*

entity from the project owner, established in the European Parliament according to the decision of the Bureau will be consulted once operational (Note from the Secretary-General of 5 March 2024).

• Does the involvement of these third parties go beyond the development phase? *All sections from this assessment will be proposed.*

• Did you organise risk training and, if so, does this inform about the potential legal framework applicable to the AI system? *The legal frameworks to be applied are:*

- *Council Regulation concerning the opening to the public of the historical archives - Regulation (EEC, Euratom) No 354/1983;*
- *Council Regulation amending Regulation (EEC, Euratom) No 1700/2003 - Regulation (EC, Euratom) No 1700/2003;*
- *Council Regulation amending Regulation (EEC, Euratom) No 2015/496 - Regulation (EU) No 2015/496;*
- *The Bureau decision of 2 July 2012 on rules on document management in the European Parliament established the framework for an effective and comprehensive document management policy for the Institution;*
- *Regulation 2018/1725 for the rules applicable to the processing of personal data by European Union institutions, bodies, offices and agencies.*

• Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas? *This responsibility is to be estbalished in the context in the European Parliament according to the decision of the Bureau (Note from the Secretary-General of 5 March 2024).*

• Does this process include identification and documentation of conflicts between the 6 aforementioned requirements or between different ethical principles and explanation of the 'trade-off' made? *This document will evolve according the recommendations applicable to this domain according the AI guidelines in the European Parliament.*

• Did you provide appropriate training to those involved in such a process and does this also cover the legal framework applicable to the AI system? *This training approach is to be established in the context in the European Parliament according to the decision of the Bureau (Note from the Secretary-General of 5 March 2024).*

• Does this process foster revision of the risk management process? [redacted]

• For applications that can adversely affect individuals, have redress by design mechanisms been put in place? *In case such affects would be reported and proven, the appropriate adjustments would have to be implemented to mitigate this affect.*