

**Fang Chen** and **Jianlong Zhou**

October 2020

Non-Biased Talent Shortlisting Algorithm:

Ethical AI Validation Framework

UTS CRICOS 00099F

School of   
Communication

UTS CRICOS 00099F

* 1. Questionnaires

Checklist style questionnaires for principles are set up to validate each ethical principle. Figure 1 shows the ethical principles and their segments used for questionnaires in the ethics validation framework. Table 1 shows questionnaires used to validate ethical principles.

A close up of text on a white background

Description automatically generated

Figure . Ethical principles and their segments used for questionnaires.

Table 1 Questionnaires used to validate ethical principles.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ethical Principles** | **Principle Segments** | **Checklist Questions**   * Sub-questions   + Question answer prompts/examples | **Question Type** |
| **Trans-**  **parency** | **Traceability** | Does the supplier document the information on the method of training the algorithm, including which input data was gathered and selected, and how this occurred?   * Are there documents recording the information on the method of training the algorithm?   + Yes/No   + Accessible through a link: uts.edu.au   + By request. * Which input data was used for training?   + What is the data source; * How the data was gathered?   + The data is public and free;   + The data is public but needs purchase;   + The data is non-public;   + User provides data; * How the data was split for training?   + Temporal cross validation M-N;   + Random cross validation M-N; * How the data was preprocessed for training.   + How is domain knowledge integrated into features used;   + Data cleaning – missing, duplicated, invalid values;   + Data consistency – match time, location, id, key… | Qualitative |
| Does the Supplier document the information about the data used to test and validate?   * Is there a document recording the information about the data used to test and validate?   + Yes   + Contents include data source;   + Integrate domain knowledge into features used;   + Important factors;   + Accessible through a link or by request; * How the data was gathered?   + The data is public and free;   + The data is public but needs purchase;   + The data is non-public;   + Customer provides data; * How the data was selected?   + Cross Validation; * What test and validation approaches were used?   + Validation data (another dataset, synthetic dataset, how many datasets)   + Measurement – accuaacy, F1…   + Basline methods   + Results   + Parameter tunning | Qualitative |
| Does the supplier document the outcomes of or decisions taken by the algorithm?   * Is there a document recording the outcomes of or decisions taken by the algorithm?   + Yes   + Accessible through a link or by request * How were outcomes or decisions recorded?   + In testing time order and information include data used, parameters, outcomes/decisions   + In the algorithm version oder and information include data used, parameters, outcomes/decisions | Qualitative |
| **Communi-cation** | Does the supplier clearly communicate characteristics, limitations, risks and potential shortcomings of the AI system?   * Does the supplier communicate the characteristics of the AI system?   + Yes * How the characteristics were communicated;   + Doc/report   + Demo   + Accessible through a link or by request; * What are the characteristics?   + Better performance;   + Easy to use   + Flexibility, effective usage * Does the supplier communicate the limitation of the AI system?   + Yes * How the limitations were communicated;   + Doc/report   + Accessible through a link or by request; * What are limitations;   + Need big data   + Statistically sound * Does the supplier communicate the risk of the AI system?   + Yes * How the risks were communicated;   + Doc/report   + Accessible through a link or by request; * What are risks;   + Wrong prediction for some data | Qualitative |
| **Explainability** | Does the supplier provide an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?   * Is there an explanation provided on system outputs;   + Yes * What are the explanation methods;   + Feature importance; * How users can access explanations.   + Picture/doc/by request | Qualitative |
| **Faithfulness of explanation**: Are relevance scores in explanations indicative of "true" importance? | Quantitative (Algorithm) |
| **Monotonicity of explanation:** Whether adding more positive evidence increases the probability of classification in the specified class? | Quantitative (Algorithm) |
| **Fairness** | **Bias avoidance** | **Data bias:** Does the supplier use the un-biased data? | Quantitative (Algorithm) |
| **Model bias:** Does the supplier establish a strategy or a set of procedures to avoid creating or reinforcing bias in the AI system? (Whether the AI system provides a set of procedures to avoid creating or reinforcing unfair bias?) | Quantitative (Algorithm) |
| **Model stability**: Does the supplier put in place processes to test and monitor for potential biases during the development and deployment phase of the system? (Whether the bias solutions are stable for different data?) | Quantitative (Algorithm) |
| **Accessibility and universal design** | Does the supplier involve or consult the community during the development phase of the AI system?   * Is the community involved or consulted during the development phase of the AI system?   + Yes * Is the community’s involvement or consultation documented?   + Yes   + Accessible through a link or by request; * Who are involved or consulted;   + Community person;   + Standardised documents, e.g. papers, reports; * What are comments from the community.   + Comments from the community; | Qualitative |
| Does the supplier assess whether the customer team involved in building the AI system is representative of target customer audience?   * Is the customer team involved in modelling/ developing/ training/ assessing the AI system?   + Yes. The customer is asked to nominate a team to assess the outputs of the AI outputs. * How is the customer team membership arrived at (with relation to representation and diversity)?   + The teams membership is assessed to identify level of representation of elements including: age, gender, ethnity, role, disability; * How does the customer team assess and respond to the AI system in this phase?   + “Fit-for-purpose” and functional assessment.   + Individual view of the solution from the perspective of the missing representation.   + This feedback is considered against that received from other cohorts. | Qualitative |
| Is the initial customer team representative of the wider population, considering also of other groups who might tangentially be impacted?   * Is people beyond direct customer who minght tangentially be affected be involved in the testing?   + People that decisions from AI are involved. * How are people beyond direct customer involved?   + Provide expectations at the design stage;   + Test AI solutions at the testing stage. | Qualitative |
| **Stakeholder participation** | Does the supplier consider a mechanism to include the participation of different stakeholders in the AI system’s development and use?   * Are different stakeholders included in the AI system’s development and use?   + Yes * Are the involvement activities of different stakeholder documented?   + Yes   + Accessible through a link or by request; * Who are stakeholders involved?   + The customer;   + Parties related to the customer; * What are their activities?   + Provide the requirements;   + Suggest the usabilities; * How is user feedback sought?   + Kickoff activities, participantion brief. | Qualitative |
| **Accou-ntability** | **Auditability** | Does the supplier establish mechanisms that facilitate the system’s auditability, such as ensuring traceability and logging of the AI system’s processes and outcomes?   * Is there a mechanism/s that facilitates audit of the AI systems processes and outcomes?   + Yes * What are these mechanisms, and which systems do they log?   + The mechanism includes a detailed logging of to trace AI system’s processes and outcomes. It includes...   + List of items traced and logged: Training, testing; Data; Outcomes;   + Automatic/Manual logging * How do they link to algorithm versions and data sets?   + Version control system is used. * How does a user revisit a historic decision?   + Provide an interface for visiting a historic decision. | Qualitative |
| **Minimising and reporting negative impact** | Does the supplier carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected?   * Has a formal impact assessment been done on the AI system?   + Yes. This AI solution is subject to a standard/guide. We have done a formal impact assessment following it. * Which stakeholders are considered in the assessment?   + Customer users   + Customer management   + Customer employees   + Customer contractors * How are key risks assessed?   + We follow the guide/standard of xx to assess risks. * What are the results of the assessment;   + Low risks found | Qualitative |
| Does the supplier establish processes for users to report potential vulnerabilities, risks or biases in the AI system?   * Can a user report potential vulnerabilities, risks or biases in the AI system?   + Yes. The features of potential vulnerabilities, risks or biases in the AI system are highlighted and allows users monitor those features. * How can a user do this?   + The user interface of the AI solution provides “report an issue” feature.   + Issues are identified in two ways as shown below. | Qualitative |
| **Documenting trade-offs** | Does the supplier establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?   * Does the supplier document “trade-offs” inherent in the approaches taken to address problems AI solved?   + Yes * Please identify deliberate and identified (as a side-effect) “trade-offs” and summarise impact and justification.   + Detailed trade-offs | Qualitative |
| **Ability to redress** | Does the supplier establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?   * Does the supplier have mechanisms to allow redress following cases of harm or adverse impact?   + Yes * What instances of harm/adverse impact can be redressed, and what mechanism for each?   + Incorrect/incomplete data from customer systems or public platforms. * How does the mechanism provide redress?   + Describe the mechanism. * How does the mechanism prevent from recurring?   + Update model regularly | Qualitative |
| **Privacy** | **Respect for privacy and data protection** | Does the AI system take measures to enhance privacy, such as via encryption, anonymisation and aggregation?   * Does the AI take measures to enhance privacy?   + Yes * What measures does the AI take to enhance privacy?   + Right to use   + Consider appropriate/acceptable use   + Authorised and authenticated access   + Encryption, anonymisation or aggregation * How were the measures determined to be adequate and sufficient?   + The measures implemented are based on the Australian Privacy Principles | Qualitative |
| Does the supplier build in mechanisms for notice and control over personal data in use cases (such as valid consent and possibility to revoke, when applicable)?   * Does the supplier hold consent from the individuals for the storage and processing of their personal data?   + Yes * What is the nature and scope of any consent held?   + The consent will be an electronic confirmation from the individual that they accept the terms and conditions. * How is consent captured for an individual?   + We contact the individual and seeks from them confirmation that they accept the customers terms. The record of this electronic acceptance is recorded in our database. * Can the individual seek to have their data removed from the system?   + Any request received by the customer to have an individual’s data removed from the system is passed on to our system and actioned within 24 hours. | Qualitative |
| **Quality and integrity of data** | Does the supplier align the AI system with relevant standards (for example ISO, IEEE, GDPR) or widely adopted protocols for daily data management and governance?   * Are there the relevant standards or protocols used?   + Yes * What standards or protocols were used?   + IEEE Standardards ( which specific standards?)   + GDPR   + Others * What measures were taken to make sure standards or protocols were aligned?   + Company sets up a work pipeline following on IEEE standards | Qualitative |
| Does the supplier put in place processes to ensure the quality and integrity of data? How is the supplier verifying that data sets have not been compromised or hacked?   * Does a data quality and integrity process exist?   + Yes * Does the process check against an external source to ensure the quality and integrity of data?   + Yes   + The external source is Linkedin data * What source of truth is used to establish ongoing consistency?   + Provide the source of truth. * How is consistency confirmed?   + Provide approaches for the consistency confirmation. * How are inconsistencies managed?   + Provide approaches to manage inconsistencies. * How do we verify that a data set has not been compromised or hacked?   + Provide approaches to verify a data not being compromised or hacked. | Qualitative |
| **Access to data** | Does the supplier ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?   * Does an access oversight mechanism exist?   + Yes * What information is logged?   + Describe information logged. * How are the logs managed (created, stored) and how do invalid accesses get identified (notification or review, to/by whom)?   + Describe approaches used for the management of logs and approaches to identify invalid accesses. * How is access to data withdrawn (staff or system off-boarding)?   + Describe approaches used to withdraw access to data. | Qualitative |
| Does the supplier have the process and/or infrastructure to restrict access to user data?   * Do processes/infrastructure to restrict access to user data exist?   + Yes * What are the processes/infrastructure implemented to restrict access to user data?   + Every customer’s data is stored separately and encrypted at rest.   + Secure protocols are used to connect with customer systems using Transport Layer Security 1.2 for HTTPS encryption. * How do the implemented processes/ infrastructure restrict access to user data?   + Access to customer data is only provided to select employees to troubleshoot in the event of a customer issue that needs to be resolved. Arbitrary access is prohibited, and every access is logged | Qualitative |