The Alan Turing Institute

Centre for Emerging Technology and Security

Structured Framework for Assessing Proportionality of Privacy Intrusion of Automated Analytics

Factors when considering intrusiveness of automated analytic methods

1. Datasets

1.1 Datasets for analysis

- What are the datasets?
- What are the sensitivities and are there any data types of concern e.g. legally privileged?
- What are the granularities and is it possible to infer detailed information regarding specific individuals?
- Are there any issues in terms of collateral intrusion?
- Are there concerns related to integrity?
- Could the volume be reduced?

1.2 Datasets for training (questions only applicable for AI methods)

- What are the datasets and do the volumes meet requirements?
- What are the sensitivities and are there any data types of concern e.g. legally privileged?
- What are the granularities and is it possible to infer detailed information regarding specific individuals?
- Are there any issues in terms of collateral intrusion?
- Is the data consistent with the data for analysis in its granularity, integrity, and biases and if not, what are mitigations?
- What is the quality of data labelling (if applicable)?
- Can the purpose be achieved using anonymised or synthetic data (as opposed to real citizen data)?
- If real citizen data is used, could the volume be reduced?

1.3 Datasets for testing

- What are the datasets and do the volumes meet requirements (i.e., yield statistically meaningful results)?

Ardi Janjeva, Muffy Calder and Marion Oswald, "Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analytics," *CETaS Research Reports* (May 2023).

- What are the sensitivities?
- Are there any issues in terms of collateral intrusion?
- Is the data consistent with the data for analysis in its granularity, integrity, and biases and if not, what are mitigations?
- Is the data consistent with the training data in its granularity, integrity, and biases and if not, what are mitigations?
- Can the purpose be achieved using anonymised or synthetic data, as opposed to real citizen data?
- If real citizen data is used, could the volume be reduced?

2. Results

- Are results produced regularly or upon request?
- Is this a step change in the scale of results that can be generated?
- Does the analysis have the potential to result in further data being collected?
- What is the granularity, accuracy of prediction, and explainability of results?
- Will they support a specific/discrete investigation or a strategic/general purpose solution?
- Will they alone be used to inform subsequent decision making? If not, what other factors will be taken into account?
- Who/which systems require or will have access to results, or reports based on the results? Is there automatic chaining of analytics?
- How does the intrusion scale from individuals to different populations? For example, is it constant, additive, or multiplicative?
- How could the intrusion affect communities with protected or sensitive characteristics?
- Could any inaccuracies, as a consequence of bias, uncertainties, or mismatch between training, testing, and analysis data, lead to adverse outcomes?
- What are the error reporting processes in the event of adverse outcomes or misdirected actions deriving from these results and subsequent decisions?

3. Human inspection

- When is human inspection of intermediate results expected to occur and at which points? For example, does it happen after one or several automated filtering steps?
- To what extent are such intermediate results understandable by a human and potentially actionable?
- Does any automated filtering inform another decision or automated system (before human inspection)?

- Are levels of uncertainty in the algorithms known and if so, what is their impact on the volume and sensitivity of data requiring human inspection?
- Is the expected human inspection feasible? For example, is there an upper bound on the amount of data that can be reviewed and number of people required?

4. Tool design

- What biases and constraints exist within the algorithm design (excluding training data), including assumptions about data for analysis and uncertainties of prediction?
- What is the amount of data required to train and test model(s), is it minimal?
- What is the expected lifetime before retraining is required and how is performance monitored?
- How much control does the user have over thresholds within the algorithm(s) (if applicable)?
- Is this automating a new capability or an existing process?
- Is this an established or a prototype tool and is it novel in this domain, business as usual in this domain, or business as usual in another domain?

5. Data management

- Who/which systems have access to the datasets and are you assured there is suitable access control?
- What are the retention and deletion policies for all the datasets and is the training data extended retention policy consistent with the retraining lifecycle and any requirement to revert to a previous model?
- Are you assured by whoever is retaining the data that it is protected from loss and corruption?

6. Timelines and resources

- What is the urgency, gravity, and extent of potential harm?
- What are the timescales for each of the steps and is there any flexibility?
- Are there adequate computational resources (processing power, storage) and training data suitably labelled (if required), to meet those timescales?

This framework was developed as part of a CETaS research project exploring privacy intrusion arising from the use of automated analytics within the national security context. For more information, please read the full report ***Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analytics*** *on the CETaS website.*

Ardi Janjeva, Muffy Calder and Marion Oswald, "Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analytics," *CETaS Research Reports* (May 2023).