**USE CASE 1: ROXANNE**

**Pre-reading:**

<https://www.forbes.com/sites/aparnadhinakaran/2021/09/10/overcoming-ais-transparency-paradox/>

**Introductory narrative**

Transparency in public sector ML - approaches in TRI research projects

ROXANNE – Real-time network, text and speaker analytics for combating organised crime.

The ROXANNE project, funded by the European Commission, seeks to support law enforcement authorities in creating tools to assist in the discovery of criminal networks and the identification of their members, but using speech and language analysis technologies, visual analysis and network analysis. ROXANNE tools are intended to enhance criminal network analysis by providing a framework for extracting intelligence from speech, language and video sources and automating time consuming tasks in processing evidence of this nature. In ROXANNE, TRI is leading an a legal and ethical analysis, and which includes integration with INTERPOL and EU legal and ethical frameworks.

Work on transparency in ROXANNE focused around the importance of law-enforcement end-users being able to understand how the tools work, it being particularly important for investigators to be able to explain their actions and potentially their use of ROXANNE tools in a court of law. As part of its ethical and legal analysis TRI developed a set of questions for the partner organisations in the project that were developing the ML-based tools. These questions dug into the transparency and bias issues that tools of this type can often encounter. It was a useful exercise for us to understand the extent to which the technology developers and law enforcement agencies in the project understood transparency, their transparency needs, and what measures they felt able to put in place. The questions TRI asked were:

**Questions to pose to the audience**

* How is traceability implemented (e.g., logging)? What measures have been taken to ensure the tool is auditable? How do you detect when things go wrong?
* How is the processing in this tool understandable to technical experts? Are there any differences for use, validation, and evaluation of tools? Has this changed during the project?
* How will end-users be able to interpret the results of this tool? How will they understand the rationale for the tool results/outputs?
* If conspicuous (i.e., unusual) activity could be highlighted by your tool, have you considered whether thresholds need to be evaluated to ensure innocent people are given the benefit of doubt?
* Does the documentation/training manual explain to end-users how the tool works and how to understand outputs? Does it respond to all the points in the training guidance document provided by Trilateral?

**Further/follow-up narrative**

The ML technologies developed by the technical partners in ROXANNE are not inherently interpretable, and to an extent are black-boxes in needs of post-hoc explanations. As such, much of the work on the practical implementation of transparency was shifted towards the part of the project that dealt with end-user training. Trilateral developed a series of prompts for technical partners to consider when writing training materials so that they could consider additional information to improve the information included in the training materials that could be relevant to transparency.

A consistent safeguard across the ROXANNE technologies was implementing logging of use and outputs of the tools (to support longer term accountability and audit). Several tools implement the display of confidence levels or percentages to assist users.

In collaboration with Interpol, ROXANNE circulated a second larger external survey to Interpol member authorities. While the main aim of the study was to identifying the appropriate legal frameworks surrounding LEA use of machine-learning powered digital forensics tools, a key additional finding for us from this survey was that for most responding police forces, the outputs from digital forensic tools were largely represented in court through statements from a police investigator, rather than interacted with directly.