Robotics & Automation Engineering Journal
ISSN: 2577-2899

Juniper Publishers
key to the Researchers

Research Article
Volume 7 Issue 1 - December 2025
DOI: 10.19080/RAEJ.2025.07.555702

Robot Autom Eng J

Identifying the Possibility of Indirect Leaks in AI systems

Konyavskiy V*, Agapitov D

Moscow Institute of Physics and Technology, Moscow, Russia

Submission: December 04, 2025; **Published:** December 10, 2025

***Corresponding author:** Konyavskiy V, Moscow Institute of Physics and Technology (National Research University), Moscow, Russia

Annotation

The trustworthiness of artificial intelligence systems is determined, in particular, by the implementation of a set of technical information security measures, which must be reinforced with protection against new types of attacks aimed at extracting protected data from collections of reports (outputs). Leaks associated with this type of attack are identified as "indirect leaks." For AI systems, determining the possibility of indirect data leakage is a relevant scientific problem. To detect the possibility of indirect data leaks, a new method for analyzing sets of queries is proposed, based on the use of Jacobian matrices.

Problem statement

We consider the use of external (acquired) data. Here, to improve model quality, we use data collected by different operators such as banks, retailers, telecom operators, insurance companies, etc. Thus, the data is first enriched (in our case-by merging datasets from different operators) and then used in model development process.

At each of these stages there are specific features of working with datasets. An Enriched Dataset (ED) is formed by combining the existing datasets (D) of various operators. During model development, the developer's access to data may be restricted by classical information-protection mechanisms. However, this does not exclude the possibility of computing protected data (and/or gaining access to them) by other means. Therefore, **it is necessary to determine whether an indirect data leak from the ED is possible during model development**.

Application of Jacobian matrices and their extension to detect possible indirect data leaks

We assume the use of external data accumulated by k data Operators (DOs). Let's denote them as $DO_i, i = 1, k$. DO_i accumulates a dataset $D_i, 1, k$. Clearly, over time each DO_i grows constantly as relationships with new data subjects appear, thus, the size of the dataset does not decrease. If D_i contains p_i features for each of m_i subjects, then D_i obviously contains $d_i = m_i.p_i$ data items.

At the beginning it is necessary to form an ED. Let, for definiteness DS_1, be called the dataset being enriched. As a result of merging DS_1 with $DS_i, i = 2, k$, we obtain an ED, which contains $d \le \sum_{i=1}^{k} m_i$ data items. It's clear, that $m \le \sum_{i=1}^{k} m_i$, $p \le \sum_{i=1}^{k} p_i$ and $m << p$. Hereis the number of rows (subjects) in the ED, and p is the number of columns (features) in the ED.

Note that enrichment is called horizontal when $p > p_1$, and vertical when $m > m_1$. In both cases, ED contains $d = m$. p data items. Denote the entire collection of these data by $X = \{x^i\}, i = 1, d$.

Next comes model selection and training. The model developer chooses the best model from a family of standard models using the ED. Machine-learning technologies are well described in the literature . We only note that model quality is evaluated by computing M values of a finite set of performance metrics . Denote the vector of metric values by $Y = \{y^i\}, i = 1, M$

To compute these values, M so-called IT-pipelines are used - understood as a fixed sequence of operations on data (in accordance with [1]). In essence, this overall transformation can be described by a set of functions. Let's introduce the notation for these functions: $F = \{f^i\}, i = 1, M$.

Using the introduced notation, the goal of model developer can be described as follows: evaluate the quality of the model based on estimates:

$$Y = F(X) *$$

This is a formalization of standard practice. From the viewpoint of information security in terms of indirect leaks, we pose the following problem:

Is it possible to determine the value of $x_j^i, x^i \in X$ for some

$i \in \{1, d\}, j \in \{1, p\}$, given F and Y?

In the form of (*), this problem can be reduced [12] to computing the rank of a Jacobian matrix.

$$y_1 = f_1(x_1, x_2, ..., x_n)$$
$$y_2 = f_2(x_1, x_2, ..., x_n) \quad (1)$$
$$...$$
$$y_m = f_m(x_1, x_2, ..., x_n)$$

Where $x_i, i = \overline{1, n}$ are the parameters of a subject of interest $x \in X$ from the dataset.

Suppose the value of function from (1), say y_i, is uniquely determined by the values of the remaining functions:

$$y_i = \Phi(y_1, y_2, ..., y_{i-1}, y_{i+1}, y_m) \quad (2)$$

In this case, it is said that the function y_i depends on the others, and the functions in (1) are called dependent. Otherwise, the functions in (1) are called independent on the domain under consideration.

Now consider the Jacobian matrix composed of the partial derivatives of these functions with respect to all independent variables:

$$J = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} \frac{\partial y_1}{\partial x_2} \cdots \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} \frac{\partial y_2}{\partial x_2} \cdots \frac{\partial y_2}{\partial x_n} \\ \frac{\partial y_m}{\partial x_1} \frac{\partial y_m}{\partial x_2} \cdots \frac{\partial y_m}{\partial x_n} \end{pmatrix} \quad (3)$$

The rank $rk(J)$ of the Jacobian matrix J characterizes the independence of the functions; specifically, if $rk(J) = m$, then the functions in (1) are independent [2].

We can now determine whether an indirect leak of the parameter value is possible when applying the IT-pipelines described by (1). To do this, we add to (1) a function corresponding to a trivial pipeline - denote it by $y_{m+1} = x_i$.

Now we construct the matrix (3) for the extended set of IT pipelines, and compute the rank of the extended Jacobian matrix J^+. If $rk(J^+) = m + 1$, then all functions are independent, and an indirect leak is excluded. If, however, $rk(J^+) = m$, then the added trivial IT-pipeline $y_{m+1} = x_i$ belongs to the set of dependent IT-pipelines, and we must conclude that an indirect leak of x_i is possible. In other words, there exists a function such that $x_i = \Phi(y_1, y_2, ..., y_m)$, thus, the value of x_i is computable.

We perform these steps for each value $i = \overline{1, n}$, thereby checking for the presence of an indirect leak for every x_i.

Thus, using this method, for any set of IT-pipelines whose corresponding functions are continuously differentiable, one can determine whether an indirect leak is possible - thereby solving the problem of detecting indirect leaks for such IT-pipelines.

After detecting a potential indirect leak, one can either apply a noise-injection mechanism to the outputs of those IT-pipelines that permit the detected leak, or simply block execution of that set of IT-pipelines - thus providing protection against the exploitation of indirect leaks.

If, for some reason, it is impossible to eliminate the dependence between the functions constituting the IT-pipelines, mechanisms of differential privacy can be applied to the corresponding report data.

Conclusion

The introduced notion of "indirect leak" reveals specific properties of AI systems that must be taken into account when building trusted systems. Based on the general theory of functional dependence, a new analysis method using extendable Jacobian matrices is proposed to study the possibility of indirect leaks. The proposed method can be extended to work with metrics (IT-pipelines) whose associated functions are, in the general case, not continuously differentiable over their entire domain. This solves the posed problem of detecting the possibility of indirect data leaks in AI systems.

References

1. Konyavsky VA, Konyavskaya-Schastnaya SV, Ross GV, Raigorodskiy AM, Trenin SA, et al. (2024) "Blind" processing technology for external data in machine learning systems. Voprosy zashchity informatsii [Information Security Issues]. Moscow 2: 17-32.

2. Fikhtengolts GM (1997) Course of Differential and Integral Calculus. Saint Petersburg: Lan 1.

How to cite this article: Konyavskiy V, Agapitov D. Identifying the Possibility of Indirect Leaks in AI systems. Robot Autom Eng J. 2025; 7(1): 555702. DOI:10.19080/RAEJ.2025.07.555702

DOI: 10.19080/RAEJ.2025.07.555702

How to cite this article: Konyavskiy V, Agapitov D. Identifying the Possibility of Indirect Leaks in AI systems. Robot Autom Eng J. 2025; 7(1): 555702. DOI:10.19080/RAEJ.2025.07.555702