



digital
rights

new technologies for criminal law

Note 1 – AI Risk Matrix

The page references provided in these notes refer to the official pagination of Deliverable D3.3 as submitted to the European Commission in the context of the DIGITAL RIGHTS project.



Co-funded by the
European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Core risks

Risk of Epistemic Opacity and Lack of Explainability

Neural network models typically involve complex mathematical operations across thousands or millions of parameters. As a result, it may be difficult for judges, prosecutors, defence lawyers or defendants to understand how a specific output was generated. (p. 40)

Where algorithmic systems operate as “black boxes”, the defence may lack the technical capacity or access necessary to meaningfully contest the evidentiary value of algorithmically derived conclusions. (p. 40)

Risk of Algorithmic Error and Hallucination

AI systems generate outputs by predicting likely token sequences rather than retrieving verified facts, they may produce plausible but incorrect results, a phenomenon commonly described as “hallucination”. (p. 40)

Incorrect classifications or false correlations may influence investigative decisions or prosecutorial strategies. (p. 41)

Risk of Structural Bias in Training Data

AI models learn statistical patterns from the datasets used during their training phase. If these datasets contain biases...the resulting system may reproduce or amplify these biases. (p. 41)

Bias embedded in training data can lead to discriminatory outcomes affecting specific demographic groups. (p. 41)

Risk of Overreliance on Algorithmic Outputs (Automation Bias)

Decision-makers may place excessive trust in algorithmic outputs simply because they are generated by advanced technological systems. (p. 41–42)

Algorithmic results may be accepted without adequate scrutiny. (p. 42)

Critères d'évaluation (Risk Matrix)

Risks are classified according to:

- Normative gap
- Structural weakness
- Technological trigger
- Likelihood
- Impact

Mitigation feasibility (p. 23)

Classification des risques

Using this matrix, risks are classified into four levels:

- Low risk
- Medium risk
- High risk

Critical risk (p. 27–28)

Logique de la matrice

The matrix allows risks to be classified according to their overall severity, facilitating prioritisation when designing mitigation measures. (p. 27)



digital
rights

new technologies for criminal law

DIGITAL RIGHTS

Project

New technologies for criminal law

Note 1 – AI Risk Matrix

This content consists of structured extracts from Deliverable
D3.3 – Risk Assessment & Mitigation Plan.

Project partners



CORTE DI APPELLO DI VENEZIA



EPLO
European Public Law
Organization

