TOWARDS AN EU REGULATORY FRAMEWORK FOR AI EXPLAINABILITY

KEY TAKEAWAYS NOTE FROM CERRE EVENT

22 DECEMBER 2022

EVENT INFORMATION PAGE AVAILABLE HERE AND RECORDING HERE.

The Centre on Regulation in Europe ('CERRE') webinar "Towards an EU Regulatory Framework for AI Explainability" addressed the concept of Artificial Intelligence ('AI') explainability as it appears in current and future European Union ('EU') legislation. CERRE Research Fellow Winston Maxwell, Télécom Paris Institut Polytechnique, led the discussion, with Lucilla Sioli, European Commission, Director: AI & Digital Industry, DG CNECT; Carlos Romero Dupla, Telecomms & Media Attaché at the Spanish Permanent Representation to the EU; and Daniel Okma, Data Protection Manager, Uber.

Below are the main takeaways:

"Global" explainability is well covered by the proposed AI Act. Currently, the proposed AI Act[1] focuses on "global" explainability (that is to say, transparency obligations that permit a user to understand the overall operation of the algorithm, its weaknesses, how it was developed and trained, and its approved use environments). Annex IV of the proposed AI Act requires extensive information on "global" explainability for high-risk AI systems, much like a warning notice accompanying a medicine. The proposed AI Act imposes technical documentation, user notification, and human control requirements that ensure global explainability.

"Local" explainability is absent from the proposed AI Act. "Local" explainability refers to the ability of a user, citizen, or data scientist, to understand a specific algorithmic result. For example, by being able to answer the following types of questions in their context: '*Why did the algorithm generate this risk score for me? Why was my loan denied? Why did the system misclassify this image?*' and so on. This aspect of explainability is absent from the proposed AI Act, even though it is in practice the biggest challenge for providers and users of algorithms. Data scientists are seeking ways to make the output of deep learning algorithms more explainable, training the algorithm in some cases to focus on causal elements and knowledge, rather than just on correlations. An "explanation" is, however, not always a valid "justification". An algorithm may give a lower credit score because of the kind of browser or computer a person uses. This may be an accurate explanation, but it will not constitute a valid justification.

[1] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 Final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

Local explainability as a human right. Two decisions of the European Court of Justice have imposed local explainability in the context of algorithms used to detect risks of terrorism. Human operators, who receive and analyse individual alerts, must be able to understand the reasons for the alert. According to the European Court of Justice, machine learning algorithms are unable to provide an individualised understanding of reasons.[2] Therefore, the Court has said that machine learning algorithms are off the table for anti-terror detection use cases. This raises the question of whether local explainability is a human right protected by the Charter in other situations where algorithmic decisions can lead to important consequences for individual rights and freedoms.

"Vertical" EU legislation requires disclosure of "main parameters" of algorithmic decisions. Other EU legislation and legislative proposals contain requirements of explainability, but do not always use the same terms. The P2B Regulation,[3] the Online Terrorist Content Regulation,[4] the Digital Services Act,[5] the proposed Platform Workers' Directive,[6] and the revised Consumer Protection Directive,[7] all impose some form of explanation, but the terms used are often different. The P2B Regulation, which requires disclosure of the main parameters of a ranking system, and the reasons for their relative importance so as to provide an "adequate understanding" to business users, is accompanied today by Commission Guidelines[8] which describe the "how" and "why" of explanations.

The European Commission is concerned about imposing disproportionate explainability requirements. The European Commission has not proposed across-the-board local explainability requirements for high-risk AI applications in the AI Act, because doing so would be disproportionate in some cases. There will often be tradeoffs between explainability and predictive performance, and the Commission does not want to prejudge those tradeoffs in particular use cases. The Commission would prefer to leave those issues to standards dealing with particular AI use cases.

[2] Judgment of the Court (Grand Chamber) of 21 June 2022 (request for a preliminary ruling from the *Cour constitutionnelle – Belgium)* – *Ligue des droits humains v Conseil des ministres* (Case C-817/19), (OJ) C 36, 3.2.2020.

[3] Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (Text with EEA relevance) PE/56/2019/REV/1, (OJ) L 186, 11.7.2019, p. 57–79. Available at: http://data.europa.eu/eli/reg/2019/1150/oj

[4] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance), PE/19/2021/INIT, (OJ) L 172, 17.5.2021, p. 79–109. Available at: http://data.europa.eu/eli/reg/2021/784/oj

[5] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) PE/30/2022/REV/1, (OJ) L 277, 27.10.2022, p. 1–102. Available at: http://data.europa.eu/eli/reg/2022/2065/oj

[6] Proposal for a Directive of the European Parliament and of the Council on Improving Working Conditions in Platform Work, COM/2021/762 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0762

[7] Consolidated text: Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on Consumer Rights, Amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and Repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council (Text with EEA relevance) Text with EEA relevance. Available at: http://data.europa.eu/eli/dir/2011/83/2022-05-28

[8] Commission Notice Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council 2020/C 424/01, C/2020/8579, OJ C 424, 8.12.2020, p. 1–26. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020XC1208(01)

Explainability should be “meaningful”, not a box-ticking exercise. Concern was expressed that certain legislation, such as the draft Platform Workers’ Directive,[9] might impose explainability without taking into account the context and utility of explanations for the person receiving them. Explainability should be linked to its utility in helping a particular stakeholder – a user, a citizen, or a human-overseer – achieve an objective, such as being able to evaluate the accuracy of an output, or to contest a decision. This requires an analysis of which stakeholders need explanations, and why they need them. In some cases, explanations are merely a means to verify that there has been no error. In other cases, explanations help to ensure that algorithmic decision processes are fair and respectful of human values. As with the provision of information under the General Data Protection Regulation[10] (‘GDPR’), too much information, and too many explanations, can fail both of these objectives.

GDPR, Convention 108+, and explanations. AI systems that use personal data are already subject to transparency and explainability requirements under the GDPR and the Council of Europe’s Convention 108+.[11] For entirely automatic decisions, the GDPR requires disclosure of the “logic involved” in the decision, while Convention 108+ requires disclosure of the “reasoning underpinning” the decision. In addition to these specific explainability requirements, the GDPR imposes broad transparency obligations, which cover many of the same concerns as those addressed by AI explainability – making decision processes fair and respecting the autonomy and choices of the persons affected by those decisions. Specific AI explainability requirements should avoid creating a second layer of transparency requirements where the GDPR already applies.

Key recommendations:

1. Consistent with the case law of the CJUE, the proposed AI Act should acknowledge the need for “local” explanations of algorithmic decisions in cases where significant human rights are at stake. The content and form of these local explanations should be defined by the provider and the user based on the risk analysis of the high-risk AI system. To avoid excessive “box-ticking” explanation requirements, local explanations should be designed and evaluated based on their effectiveness *vis-à-vis* humans receiving the explanation, whether the recipient is the human overseer of the system verifying algorithmic outputs, or the human affected by the algorithmic decision.

[9] Supra note 6.

[10] Consolidated text: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) Text with EEA relevance. Available at: http://data.europa.eu/eli/reg/2016/679/2016-05-04

[11] Protocol Amending the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, (CETS No. 223), Council of Europe. Available at: https://rm.coe.int/16808ac918

2. Given the multiple requirements in existing and proposed EU law relating to explanations of the "main parameters" for algorithmic decisions, the European Commission should publish guidelines addressing the "how" and "why" of these requirements in different regulations and directives on digital activities, building on the guidelines already published by the Commission for the Platform to Business Regulation.[12]

[12] Commission Notice Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council 2020/C 424/01, C/2020/8579, OJ C 424, 8.12.2020, p. 1–26. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020XC1208(01)