Algorithm Audit

Feedback on DSA Delegated Regulation

Include the normative dimension of AI auditing (with a focus on recommender systems)

In addition to Article 37 of the Digital Services Act (DSA), Delegated Regulation (DR) sets out procedures, methodologies and templates for third-party auditing of Very Large Open Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). The DR builds upon established sector-specific risk management frameworks to provide procedural guidance for AI audits. However, the regulation lacks provisions to disclose normative methodological choices that underlie AI systems (e.g., recommender systems), which is crucial for evaluating associated risks in a meaningful way (as mandated by DSA Article 34). To illustrate this limitation, we elaborate on methodological crossroads that determine the performance of recommender systems and its downstream risks. We make concrete suggestions how the definition of 'inherent risk' (Article 2), audit methodologies of risk assessments (Section IV) and the audit report template (Annex I) set out by the DR should be amended to incorporate normative dimension of AI auditing in a meaningful way. Only if both the technical and normative dimension of AI systems are thoroughly examined, risk assessed under the DSA will empower the European Union and its citizens to determine what public values should to be safeguarded in the digital world.

Suggestions for advancing the proposed Delegated Regulation (DR)

> Amend the definition of 'inherent risk' in Article 2 DR: Add the underlined clause to the definition of 'inherent risk': "…the nature, the activity, the normative design choices and the use of the audited service…".

> Include clause on the normative dimension of risk assessments in Article 13(a)(ii) DR: Add the underlined clause to the analysis obligations: "How the audited provider assessed each risk, including how it considered the probability, normative dimension and severity of the risks …".

> Include clauses on normative considerations and methodological design choices to Annex I Section D.II – Template for audit report: Add the underlined clause to question 3(a): "[…] justification of the normative choice of those procedures and methodologies

June 2023

(including, where applicable, a justification for the choices of standards, benchmarks, methodological design choices, sample size(s) and sampling method(s)):".

More details about the above suggestions can be found in the section Suggestions for advancing the proposed Delegated Regulation (DR).

AI audits: A technical and normative dimension

Traditionally, audits in the financial, medical, and IT sector are regarded as technical check-list routines. Standardized procedures ensure that records, statements, and processes undergo the same examination and evaluation regardless of the auditor involved. For instance, in the context of financial risk modelling, central banks have developed procedures to systematically quantify the impact of inaccurate mortgage default prediction models on banks' financial stability. Similarly, in the field of drug testing, regulatory agencies collaborate with subject matter experts to establish objective safety measures to manage trial risks. In short, established audits are tailored to specific sectors, technologies, and contexts to manage risks effectively.

Due to the subjective nature of risk, developing auditing methodologies for all-purpose technologies (like AI systems) pose significant challenges. Auditing a financial asset-liability model differs fundamentally from auditing a recommender system, because quantifying risk in monetary terms is less influenced by subjective values than engineering recommendation systems (see Example section). Pursuant to the DSA[1], VLOPs and VLOSEs must perform risk assessments of recommender systems (Article 34) and arrange independent annual audits to ensure compliance (Article 37). We argue that besides technical examinations that ensure robust AI engineering, such as evaluating logging protocols and model monitoring capabilities, normative aspects must get a significantly more prominent role in the risk definition (Article 2), audit methodologies of risk assessment (Section IV) and audit report templates (Annex I) set out by the delegated regulation (DR)[2]. In order to provide tangible suggestions, we first elaborate on methodological choices underlying one specific type of all-purpose AI system, i.e., recommender systems.

Example – Methodological choices guiding recommender systems

Recommender systems play a crucial role in many digital services provided by VLOPs and VLOSEs, such as advertising, search and news feeds. The goal of recommender systems is to provide suggestions most pertinent to a particular user – a process commonly referred to as *learning to rank* (LTR). LTR methods rely on feature engineering, in which characteristics are learnt from user-system interaction, essential for both personalized and non-

[1] DSA: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825

[2] Delegated regulation https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13626-Digital-Services-Act-conducting-independent-audits_en

personalized recommender systems. When developing LTR systems, there are several methodological choices that must be carefully considered due to its impact on downstream risks. We highlight two methodological crossroads:

> Statistically biased or unbiased feature extraction – A first methodological choice concerns the use of statistically biased or unbiased feature extraction methods. Statistically biased methods operate under the assumption that clicks occur independently of the position of an item. Examples of widely used biased LTR methods are RankNet, ListNet, LambdaMART and RankSVM. Unbiased methods challenge the assumption of independent clicks and adjust rankings based on click propensities. Assumptions regarding click behaviour have implications for ranking performance and downstream risks associated with recommender systems, including confirmation bias (reinforcing existing believes or preferences) and popularity bias (promoting popular items and content that is relevant for other users, leading to positive feedback loops). So, risks assessment of recommender system must include a review of rationales why a certain feature extraction method is chosen.
> Evaluation metrics for machine learning – A second methodological choice concerns evaluation metrics for machine-learned ranking. For instance, a *pairwise approach* to assign relevance scores to suggestions treats LTR as a classification task. Alternatively, a *listwise approach* directly optimizes suggestions according to a specific evaluation metric, e.g., Mean and Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), Precision@n. These methodological choices form the foundation for evaluating the performance of recommender systems. Downstream risks, such as confirmation and popularity bias, should therefore be evaluated and documented considering various evaluation metrics and varying feature extraction methods.

Suggestions for advancing the proposed Delegated Regulation (DR)

Amend the definition of 'inherent risk' in Article 2 DR – Definitions

Article 9 of the DR specifies that "audit risk analysis shall consider *inherent risk*, *control risk* and *detection risk*". 'Inherent risk' is however vaguely defined in Article 2 of the DR. More specific guidance should be provided how risks relating to subjective concepts, such as "…the nature, the activity and the use of the audited service", can be assessed. Building upon the above recommender system example, we argue that the methodological choices that guide AI systems are underrepresented in the current definition of 'inherent risk'. We suggest amending this definition such that the intrinsic normative dimension that influences risks of AI systems is incorporated. We suggest to include the underlined clause to the definition of 'inherent risk': "…the nature, the activity<u>, the normative design choices</u> and the use of the audited service…".

Include clause on normative dimension of risk assessment in Article 13(a)(ii) DR

Pursuant to Article 34(2)(a) under the DSA, Article 13(a) DR prescribes that a risk assessment of the "design of recommender systems" should include an analysis "whether the audited provider has diligently identified, analysed, and assessed the systemic risks". Building upon the above recommender system example, we argue that this risk assessment can only be conducted in a meaningful way, if the normative dimension of the system is included. We therefore propose that the following underlined clause should be added to Article 13(a)(ii) DR: "How the audited provider assessed each risk, including how it considered the probability, normative dimension and severity of the risks …".

Additions to Annex I Section D.II – Template for the audit report

Pursuant to Article 5(1)(a) of the DR, VLOPs and VLOSEs shall transmit to third-party auditing organisations "benchmarks used […] to assert or monitor compliance […], as well as supporting documentation". Building upon the above recommender system example, we argue that normative considerations that underly the selection of these benchmarks should be asked out more decisively in this phase of the audit. Choices for certain feature extraction methodologies, e.g., statistically biased or unbiased approaches or evaluation metrics, e.g., MAP, NDCG, Precision@n (see Example section), are essential to know when auditing benchmarking analysis for AI systems, including a critical review of documentation of these choices. The following underlined clause should therefore be added to Question 3(a) of Section D.1 *Audit conclusion for obligation* Subsection II. *Audit procedures and their results*: "[…] justification of the normative choice of those procedures and methodologies (including, where applicable, a justification for the choices of standards, benchmarks, methodological design choices, sample size(s) and sampling method(s))".

Work of Algorithm Audit

Audit commissions	Advising on ethical issues emerging in concrete algorithmic practices through deliberation, resulting in *algoprudence*
Technical tooling	Implementing and testing technical tools to detect and mitigate bias in data and algorithms
Advocacy	Contributing to public debate on responsible use of algorithms
Knowledge sharing	Sharing techno-ethical insights with society, policy makers, algorithm subjects and others