RESEARCH OUTPUTS / RÉSULTATS DE RECHERCHE

An Empirical Study on Compliance with Ranking Transparency in the Software Documentation of EU Online Platforms

Sovrano, Francesco; Lognoul, Michael; Bacchelli, Alberto

Link to publication

Citation for pulished version (HARVARD):
Sovrano, F, Lognoul, M & Bacchelli, A, *An Empirical Study on Compliance with Ranking Transparency in the Software Documentation of EU Online Platforms: 46th International Conference on Software Engineering: Software Engineering in Society*, 2024, Web publication/site.

An Empirical Study on Compliance with Ranking Transparency in the Software Documentation of EU Online Platforms

Francesco Sovrano*
University of Zurich
Zurich, Switzerland
francesco.sovrano@uzh.ch

Michaël Lognoul*
University of Namur (CRIDS, NADI)
Namur, Belgium
michael.lognoul@unamur.be

Alberto Bacchelli
University of Zurich
Zurich, Switzerland
bacchelli@ifi.uzh.ch

ABSTRACT

Compliance with the European Union's Platform-to-Business (P2B) Regulation helps fostering a fair, ethical and secure online environment. However, it is challenging for online platforms, and assessing their compliance can be difficult for public authorities. This is partly due to the lack of automated tools for assessing the information (e.g., software documentation) platforms provide concerning ranking transparency. Our study tackles this issue in two ways. First, we empirically evaluate the compliance of six major platforms (Amazon, Bing, Booking, Google, Tripadvisor, and Yahoo), revealing substantial differences in their documentation. Second, we introduce and test automated compliance assessment tools based on ChatGPT and information retrieval technology. These tools are evaluated against human judgments, showing promising results as reliable proxies for compliance assessments. Our findings could help enhance regulatory compliance and align with the United Nations Sustainable Development Goal 10.3, which seeks to reduce inequality, including business disparities, on these platforms.

Data and materials: https://doi.org/10.5281/zenodo.10478546.

CCS CONCEPTS

• **Software and its engineering** → **Documentation**; • **Information systems** → **Data analytics**; • **Applied computing** → **Law**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Software Documentation, EU Regulations, Compliance Assessment, Ranking Transparency, Explainability, Online platforms

LAY ABSTRACT

In the digital world, software powers online platforms such as marketplaces and search engines. For instance, software determines

*Both authors contributed equally to the paper

the display order of products on Amazon or hotels on Booking.com. Understanding these systems is vital for businesses, yet the explanations can be complex. The European Union mandates online platforms to reveal the main parameters of their ranking algorithms, aiming for fairness and reducing inequalities. However, ensuring these platforms comply is difficult due to the lack of a standard evaluation method. Our study examined explanations from Google, Yahoo, Bing, Amazon, Booking, and Tripadvisor. We gathered insights from three legal experts and over a hundred people to gauge the clarity of these explanations, finding significant variations in clarity across platforms. Reviewing these explanations manually is cumbersome. To streamline this, we developed automated tools and compared their effectiveness against human assessments. Our discussion revolves around how these tools can enhance transparency on online platforms.

1 INTRODUCTION

In today's digital landscape software is more than just lines of code, it is a driving force behind economic activities and business operations [6, 7, 15]. This impact is particularly evident in the online platforms' ecosystem, especially regarding intermediation services (e.g., marketplaces) and search engines. The software used by the providers of these types of services plays a significant role in shaping downstream business dynamics [8]. As software increasingly becomes the backbone of society, attention is shifting towards its interaction with people. One critical element of this relationship is the need for clear and transparent software documentation, often not provided by default to those who interact with software.

To address this issue, the European Union has introduced legislation such as the Platform-to-Business (P2B) Regulation (EU) 2019/1150 [12]. Among other requirements, this Regulation imposes on platforms to provide easily accessible, comprehensible, and detailed documentation explaining how ranking mechanisms work and affect sellers, including business and corporate website users (Article 5 and Recital 25 [9]). Such a measure aims to create a more equitable environment for businesses that rely on online platforms to connect with consumers. This resonates with Target 10.3 of the United Nations Sustainable Development Goals [25], which aims to ensure equal opportunity and reduce inequalities within and among countries.

However, producing documentation adhering to the P2B Regulation is challenging for online platforms, as recently pointed out by the European Commission (EC) [9]. Additionally, monitoring the quality of documentation is challenging for platforms and regulators, affecting compliance assessment, law effectiveness, and social well-being. This is largely due to the lack of standardized, automated tools for assessing compliance across different platform

providers. Hence, given the absence of standardized assessment tools, we hypothesize to encounter varying levels of compliance among platforms regarding ranking transparency.

To explore the extent to which online platforms actually produce adequate documentation, as intended by the law, we first conducted a manual assessment involving three legal experts from Belgium (with respectively 7, 6 and 5 years of experience in legal research) and more than 130 participants. Our focus is on evaluating whether the software documentation from key online intermediation services (i.e., Amazon, Tripadvisor, and Booking) and online search engines (i.e., Google, Bing, and Yahoo) genuinely aids businesses in understanding ranking mechanisms and can fulfill the objective pursued by the Regulation. We expect this empirical study to validate our initial hypothesis, affirming the variance in the quality of explanations provided by different platforms.

As a second step, we investigate whether automated tools can effectively assess compliance as a proxy for traditional surveys. Given the recent popularity of generative AI, we experiment with ChatGPT as a baseline assessment tool. We also propose an alternative approach, which combines ChatGPT with answer retrieval technologies, offering a more transparent assessment based on a novel metric, DoX [33, 34], designed to quantify the degree of explainability of documentation. We test these automated tools to evaluate their correlation with human judgments, as obtained from our empirical study. We aim to identify these systems' strengths and weaknesses in predicting compliance and explainability.

The primary contributions of this work include:

(1) A checklist derived from the P2B Regulation and European Commission guidelines for platform documentation quality assessment.
(2) A method to automate assessment using the aforementioned checklist through ChatGPT.
(3) A more deterministic, non-hallucinatory, and transparent assessment tool than ChatGPT, which is based on a theory of explanations from Ordinary Language Philosophy [1, 32] and answer retrieval technology.
(4) An in-depth analysis of ranking documentation from six major online platforms involving three legal experts and over 130 potential customers.
(5) Methods that closely align with expert views on compliance, illustrating their potential for ongoing monitoring by authorities and the public.
(6) Source code for the automated tools and referenced data, including manual assessments and checklist [30].

Considering the long-term societal ramifications, we argue that effective automated tools can lead to the development of more compliant, transparent software documentation and, consequently, more equitable online platform ecosystems for businesses. This potential transformation could significantly impact the pursuit of the UN Sustainable Development Goal 10.3, which focuses on reducing inequality.

2 RELATED WORK

This section compares our analysis of the P2B Regulation and the proposed assessment tools with existing literature.

In 2022, the Observatory on the Online Platform Economy [29] conducted a study for the European Commission, focusing on the P2B Regulation [26]. It examined terms and conditions of platforms and analyzed feedback from business users. Despite changes by major platforms like Amazon, Facebook, and Google, over half of the business users did not see improvements in the transparency of terms and conditions and ranking practices.

Subsequently (September 2023), the European Commission published a review on the implementation of Regulation (EU) 2019/1150 [9]. Their methodology included interviews and analysis of online platform terms, finding that compliance with the P2B Regulation remains a challenge despite improvements in transparency. In particular, the European Commission noted that 'in many cases, the basic information provided, for example, on ranking, is potentially insufficient' (p.9).

In contrast to the approach we propose in this paper, the review of both the European Commission [9] and the Observatory [26] employed manual assessments. They do not incorporate a detailed quantitative analysis concerning the 'quality of the explanations' found in the terms and conditions.

Additionally, Lukovic [23] examined the context and actors of platform rankings. This study details the ranking transparency obligations set by the P2B Regulation and offers critical insights into the framework's limitations. Bibal et al. [4] situate the P2B Regulation's ranking parameters within a broader EU context that mandates explanations for AI systems. This work sheds light on the intentions of EU legislators and specifies the type and extent of information essential for technical compliance. The approaches of both Lukovic [23] and Bibal et al. [4] differ from ours as they neither evaluate platforms' adherence to the Regulation nor introduce tools to enhance regulatory compliance. These elements are actually absent from existing literature, to our knowledge.

In our research, we focus on the documentation associated with the P2B regulation. In particular, we evaluate the potential advantages and limitations of leveraging ChatGPT's prompt engineering for assessment purposes. While ChatGPT has previously been harnessed for various tasks such as evaluating student responses [22] or human behavior [28], we instead apply it to assessing software documentation.

Distinguishing our work from prior studies, we propose a method designed to evaluate also the clarity and explainability of software documentation, an aspect often overlooked. This undertaking complements the growing academic interest in developing both semi-automatic and fully automatic techniques to uphold software documentation quality [16, 17].

3 BACKGROUND

This section explores the EU's P2B Regulation, fostering fairness in online platform-business interactions, alongside the European Commission's ranking transparency guidelines. It also introduces the DoX metric for measuring explainability, setting the stage for later discussions on assessment tools.

3.1 P2B Regulation & Guidelines

In 2019, the EU adopted the P2B Regulation [12] to improve fairness and transparency for online platforms in their dealings with

businesses. As noted by Prastitou-Merdi [27], it targets *online inter-mediation services*, i.e., services that allow business users to offer products to consumers and are based on contractual relationships (see Art. 2.2). The Regulation also covers *online search engines*, where users can search the internet (Art. 2.5).

The Regulation identifies two primary beneficiaries: business users and corporate website users of platform services. *Business users* are private entities acting commercially, offering goods or services to consumers via online intermediaries (Art. 2.1); *Corporate website users* are entities using an online interface, like a website or software, for professional activities to offer goods or services to consumers (Art. 2.7).

The Regulation requires online intermediation service providers to keep their terms and conditions easily accessible to business users, written in clear language (Art. 3.1). They must disclose certain information therein, such as the main criteria for ranking offers to consumers, and the reasons for their importance (Art. 5.1), as mentioned by Hacker et al. [18]. Similarly, online search engines must provide a clear, public description of the main parameters that determine ranking and their significance (Art. 5.2 [9]).

The EU legislator seeks to ensure online platforms offer explanations for business and corporate website users about ranking mechanisms [9, 29]. This includes understanding how rankings consider offer characteristics, relevance to consumers, and for search engines, website design traits (Art. 5.5). Explanations must also clarify the ability to influence ranking through payments and their impacts (Art. 5.3).

The primary objectives of the legislator in enforcing ranking transparency are twofold: *(1)* to empower businesses in the presentation of their offers online, enhancing competition in downstream markets (Recital 24) and *(2)* to enable businesses to compare ranking practices, promoting competition among online platforms in upstream markets (Recital 24). These objectives align with the United Nations Sustainable Development Goal 10.3 on reducing inequality.

To enhance compliance with ranking transparency, the Regulation required the European Commission to adopt Guidelines (Art. 5.7). In short, the European Commission's Guidelines [10] are a non-binding text (pt. 9 [10]) that elaborates on the content of the binding P2B Regulation.

The European Commission's Guidelines emphasize that online platforms' explanations should be tailored to the needs and technical abilities of average users, differing based on the service type (pt. 17). The descriptions should offer more than a mere list of main parameters and give a secondary layer of information (pt. 22). Descriptions should not be overly brief or misleading (pt. 98). Moreover, the explanations should cover all main ranking parameters, indicating exhaustive lists (pt. 24).

Regarding the form that the explanations should take, both the Regulation and the Guidelines use the term 'drafted' in English, which refers to written content.[1] To verify this, we looked into the French versions of both texts. The French translation of the Regulation refers to redacted explanations ('rédigées') within their terms and conditions for online intermediation services (Art. 3.1 and 5.1).

For search engines, the exact text uses the words 'to state' ('énoncée'), which do not imply the use of one form over another (Art. 5.2). The Guidelines, however, refer at some point to the 'redaction' of the explanations for both types of providers, potentially implying the use of a written medium in all cases (i.e., "lorsqu'ils rédigent la description [...]", pt. 99). We also looked into the Dutch version of the Regulation, which refers to 'redacted' content ('opgesteld') regardless of the type of service provider (Art. 3.1 and 5.2).

3.2 How to Measure Explainability

The Degree of eXplainability, abbreviated as DoX, is a model-agnostic metric introduced by [33, 34] to measure the clarity and depth of explanations within a text. Grounded in Achinstein's theory [1], DoX equates the act of explaining to the process of answering fundamental questions, referred to as *archetypes*. The more archetypal questions a text can answer, the higher its explainability rating. Notably, DoX can evaluate any content written in natural language once what is to be explained is defined.

In particular, DoX measures similarity, exactness, and fruitfulness, as articulated by Carnap [24]. These measurements are then merged to produce a numerical (DoX) score, providing the means for a comparison of the quality of different explanations.

To calculate DoX scores, the *relevance* of a text snippet is assessed and aggregated based on its relation to a collection of archetypal questions informed by linguistic theories (see [34]). Specifically, these relevance measurements are obtained employing pre-trained *deep language models* tailored for general-purpose answer retrieval [5, 21]. For examples on how DoX works, see the paper [34].

4 STUDY & CHECKLIST DESIGN

In this section, we detail our methodology for selecting platform documentation, which we assess later on. We also describe our method and process for developing an assessment checklist.

4.1 The Study

We independently investigate, beyond the studies and reports outlined in Section 2, the adequacy of documentation provided by several prominent online platforms as mandated by the Regulation. Our analysis focuses on software documentation from three major online intermediation services (Amazon, Tripadvisor, and Booking) and three online search engines (Google, Bing, and Yahoo).

We selected these platforms for their industry representativeness and the audience profile they serve, which is very broad and allowed us to perform a large-scale manual assessment (cf. Section 6.2).

In compliance with the P2B Regulation, intermediation services are expected to provide their users with readily accessible information within their terms and conditions (see Art. 3.1 and 5.1), and search engines within a 'publicly available description' (see Art. 5.2). However, our investigation revealed a convoluted landscape. Information on ranking, which should be easily accessible, was often dispersed across multiple pages, thus requiring users to navigate through multiple re-directions. In addition, not all the selected intermediation services consistently included this crucial data in their terms and conditions, adding to the complexity of the retrieval process.

---

[1]For definitions and examples of the transitive verb form of "draft," refer to the Cambridge Dictionary at https://dictionary.cambridge.org/fr/dictionnaire/anglais/draft.

Most platforms, however, maintained a primary web page detailing their ranking practices. This became our starting point. We confined our search to these pages and any information within one click's distance. We also observed the use of audiovisual content by some platforms, in particular by Amazon and Google, to explain ranking mechanisms. However, these videos were excluded from our study, in line with the developments held above on the provision of ranking information in a written form (see Section 3.1 and Art. 3.1, 5.1, and 5.2).

For a quantitative perspective, Table 1 presents a breakdown of the documents we retrieved from the six platforms, detailing the number of associated links and the average word count per document. Our online repository provides further details [30].

Table 1: Statistics on retrieved documents per platform.

Platform	No. of Links	Avg. Words/Doc
Amazon	5	434.40
Bing	16	964.06
Booking	7	579.42
Google	52	1679.50
Tripadvisor	10	1653.90
Yahoo	3	174.00

4.2 The Checklist

To assess the documentation quality of selected platforms, we created a checklist of essential questions, all needing positive responses for an operator to meet ranking transparency requirements. These questions were crafted by a legal expert and co-author of this study, an academic researcher with expertise in EU law on information and communication technologies and 6 years of experience. This was done using an inductive coding approach [14], allowing the checklist questions to emerge from the P2B regulation and guidelines by the European Commission. The checklist was developed based on these materials through the following steps.

Firstly, we focused on terms like 'ranking' and 'main parameter', which are legally defined and central to the P2B Regulation's explanations. However, they might be unclear to some users. Therefore, we drafted two questions to verify whether these terms were clarified in the platform's documentation (cf. Q1 and Q2).

Secondly, we incorporated each of the individual elements explicitly required by the Regulation within its own question, taking into account the wording of the legal text. These questions, therefore, address matters such as *(i)* the provision of the main parameters used for ranking (cf. Q3), *(ii)* how ranking considers the characteristics of the goods and services offered to consumers (cf. Q5), *(iii)* the extent to which such characteristics are taken into account (cf. Q6), *(iv)* the existence of payments to influence ranking (cf. Q9), and *(v)* the effects of such payments (cf. Q10).

However, the Regulation differentiates between two types of platform operators: online intermediation services and online search engines. Both are covered, but they face different requirements, as in Articles 5.1 and 5.2, and point 7 of the Guidelines. Therefore, we have created a different set of yes-or-no questions for each platform type. Importantly, this task was led by solely paying attention to the explicit requirements contained in the Regulation.

Thirdly, we drafted additional questions based on the content of the European Commission's guidelines. Whenever these guidelines explicitly raised new questions, proposed nuances to the requirements of the legal text, or gave best practices to comply with the P2B Regulation and its objectives, we designed new questions to take these elements into account and incorporated them into our checklist. These questions relate to elements like *i)* the explanation of what is most important in determining ranking (cf. Q12) and *ii)* the internal process conducted by the platforms to determine the main parameters (cf. Q14).

Fourthly, we revised certain questions' wording, without altering their meaning, to address issues in automated answer retrieval from platform documentation. For example, Q11 initially included 'all the main parameters' (cf. Guidelines, pt. 24). However, this phrasing led to issues in automated scoring, as no platform explicitly states that the provided list of parameters is exhaustive. To address this, we modified Q11 by omitting the word 'all', thereby improving the precision of automated assessments.

The list of questions for online intermediation services is:

Q1 Does the documentation explain how 'ranking' is defined/define 'ranking'?

Q2 Does the documentation explain how 'main parameter used for ranking' is defined/define 'main parameter used for ranking'?

Q3 Does the documentation provide the main parameters used for determining ranking?

Q4 Does the documentation explain why certain parameters are considered as the main ones for determining ranking instead of others?

Q5 Does the documentation explain how the ranking mechanism considers the characteristics of the goods and services offered to consumers?

Q6 Does the documentation explain the extent to which the ranking mechanism considers the characteristics of the goods and services offered to consumers?

Q7 Does the documentation explain how the ranking mechanism considers the relevance of the characteristics of the goods and services, for consumers?

Q8 Does the documentation explain the extent to which the ranking mechanism considers the relevance of the characteristics of the goods and services, for consumers?

Q9 Does the documentation explain the possibilities to influence ranking against direct or indirect payment (if any)?

Q10 Does the documentation explain the effects of payments, on ranking (if any)?

Q11 Does the documentation explain how the ranking mechanism works and, in particular, what the main parameters used are?

Q12 Does the documentation explain what is most important in determining ranking?

Q13 Does the documentation explain why specific parameters were selected as the main factors in determining the ranking of goods or services?

Q14 Does the documentation explain the internal process conducted by the provider to determine the main parameters for the ranking of goods or services?

Q15 Does the documentation explain how users can improve the ranking of their goods or services?

Q16 Does the documentation explain how users can alter the ranking of their products or services through direct or indirect payments to the provider, and what effect this has?

Q17 Does the documentation explain what the business logic behind allowing users to affect the ranking of their products or services through payments is, and what the potential consequences of this are?

The set of questions for online search engines is mainly the same but differs for three questions. Question 4 has been changed to 'Does the documentation provide the relative importance of the different main parameters used in determining ranking?'. The difference between both questions lies in two elements. Firstly, search engines do not have to provide contrastive explanations, as they are not explicitly required to explain why some main parameters were selected *as opposed to others*. Secondly, search engines have to specify the relative importance of their main parameters, whereas intermediation services do not. Moreover, in accordance with the legal text and to address the differences between types of services at stake, we added two new questions:

Q18 Does the documentation explain how the ranking mechanism considers the design characteristics of the websites?

Q19 Does the documentation explain the extent to which the ranking mechanism considers the design characteristics of the websites?

Although some of the questions, within each subset, might seem quite similar at first sight, they all call for (at least slightly) different answers, as they contain nuances that originate from the European Commission's Guidelines. For instance, questions number 4 and 13 appear to be very similar. However, question 4 puts emphasis on the contrast between the parameters selected as main ones and other parameters, whereas question 13 does not. Similarly, questions number 3 and 11 seem quite close to each other. Yet, question 3 (based on the binding text of the Regulation) only refers to the main parameters used, while question 11 (based on the non-binding guidelines) additionally considers how the ranking mechanism works in general.

5   AUTOMATED ASSESSMENT

This section delves into different strategies that use checklists to assess compliance. While conventional methods, which require manual examination of checklists against extensive documents (see Section 4), are not only lengthy but also susceptible to personal interpretations (as shown in Section 6.1), automated assessments offer regulators and public authorities significant advantages. Automated systems provide consistency in evaluation, greatly reduce the time required for assessment, and minimize the potential for human error or subjective biases. For enforcement agencies, this means faster, more reliable evaluations, ensuring more effective oversight and public trust.

Hence, we hereby discuss two computational approaches. The first uses off-the-shelf ChatGPT, while the second is based on answer retrieval technology and the DoX metric (see Section 3.2) for more robust and transparent evaluations.

5.1   ChatGPT-based Assessment

Using ChatGPT, we developed an automated algorithm to determine if documentation aligns with the criteria detailed in Section 4. Given that ChatGPT models possess inherent limitations regarding input size, our system design had to account for such constraints. For example, GPT-4 limits inputs to 8192 tokens, while the GPT-3.5-turbo-16k version allows for up to 16385 tokens.[2] Consequently, our approach breaks the software documentation into segments that adhere to GPT's token restrictions. Each section is then individually assessed against the checklist and is given a score between 1 (indicating non-adherence) and 5 (indicating complete compliance), as referenced in 6.1. After this assessment, the algorithm consolidates these scores, selecting the highest score for every checklist point.

We used the following prompt to instruct the ChatGPT model:

> *Your task is to assess the compliance of this documentation based on the following question. Conduct a compliance assessment, focusing on both the technical and legal requirements.*
> *Your assessment should start with a numerical score from 1 to 5, where 1 indicates the question is not answered at all and 5 indicates it's perfectly answered. Following the score, provide a brief explanation highlighting the strengths or weaknesses in addressing the question. Consider the completeness, clarity, and legal implications in your explanation.*
> *For example, your assessment might look like: 'Score: 3. Explanation: The question was only partially answered. While the technical aspects are covered, it lacks legal disclosures.'*
> *Question: {question}*
> *Documentation: {chunk}*

Additionally, we set the model's 'temperature' to zero to reduce randomness, thus ensuring more deterministic outcomes.

This approach showcases an elementary form of prompt engineering[3], aiming to direct the model towards expected outcomes. Yet, the foundational prompt has inherent challenges. One major issue is that segmenting extensive documentation can potentially affect the integrity of the evaluation since pertinent data might be split across various segments. The importance of this becomes evident when considering that our system employs a max operator to amalgamate scores, neglecting the differential significance of each segment. Additionally, the simplistic nature of this prompt renders the model prone to producing fabricated outputs [2, 3].

Considering these challenges, we have devised the improved strategy presented in Section 5.2. This approach incorporates advanced prompt engineering and more transparent answer retrieval mechanisms to effectively summarize lengthy documentation.

5.2   Assessing Explanation Quality with DoX

A key limitation of the standard ChatGPT-based approach presented in Section 5.1 is its inability to quantitatively assess the quality of

---

[2]The GPT-4 and 3.5 versions cited are the 0613 variants, a snapshot taken on June 13th, 2023.
[3]Prompt engineering involves formulating precise task descriptions and examples to steer the responses of language models [20].

explanations. Capturing this characteristic is quintessential to determine if the documentation provides clear explanations, especially when it comes to the legal requirements of the P2B Regulation. For instance, a platform's documentation stating 'Rankings are based on algorithms' might be accurate, but lacks the necessary depth. According to the principles of Ordinary Language Philosophy [1, 32], a proper explanation would provide more context, such as 'Rankings are determined by algorithms that consider factors like user engagement, relevance, and content quality.'

As highlighted in Section 3.2, the DoX metric is designed to assess such qualities and estimate the degree of explainability of a given documentation. Thus, we have devised a technique that combines a transparent, theory-based answer retrieval system (optimized for parsing lengthy documentation efficiently), ChatGPT (to sift through and paraphrase the retrieved answers), and the DoX metric (to estimate the explanatory power of a text).

In particular, we employ the same neural retrieval model [31] used by the DoX algorithm (cf. Section 3.2) to extract the answers to the checklist questions from the documents and ChatGPT to refine them into comprehensive explanations. Then, DoX is used to measure the explanatory quality of the answers to the checklist questions. Using the earlier example for clarity: while both answers 'Rankings are based on algorithms' and 'Rankings are determined by algorithms that consider factors like user engagement, relevance, and content quality' would achieve similar relevance ratings, the first would score lower on the DoX metric, indicating its limited explanatory depth.

Notably, a checklist is framed for binary responses, but common answer retrieval tools (such as the one we adopted) are more suited for open-ended queries, being designed for paragraph-length responses. As a result, we had to modify the checklist questions from Section 4 to accommodate a more open-ended format.

Converting these questions often means reshaping them to target the specific information needed. For instance, take the transition from the closed-ended question, 'Does the documentation explain how 'ranking' is defined?' to its open-ended counterpart, 'How is 'ranking' defined?'. While both queries focus on the interpretation of 'ranking', the second one elicits a more comprehensive answer rather than a mere yes or no.

Here is a more detailed breakdown of the steps involved in this conversion process:

(1) *Identify the Key Element*: First, identify the essential piece of information that the original question aims to find out. This is often buried in clauses like 'Does the documentation include...'.
(2) *Re-frame as Direct Question*: Next, rephrase the question to ask directly about that key element.

With these modifications, the questions become more straightforward and easier for the information retrieval system to address.

Subsequently, the answer retriever selects the top 20 best answers from documentation paragraphs. Each answer gets a pertinence score from 0 to 1, with scores near 1 indicating higher relevance to the question. On the other hand, ChatGPT identifies incorrect responses and aggregates the correct ones into a cohesive binary answer. While the answer retriever can work without ChatGPT, using them together produces better-quality answers. ChatGPT is more intelligent and better than the adopted answer retriever at detecting incorrect answers.

We used the ChatGPT version based on the GPT-4 architecture,[4] while the specific prompt template we used is:

> *Output a comprehensive answer based only and exclusively on the information within the paragraphs below (if any can be used to answer) which were extracted from the documentation to be assessed. If no paragraph can answer the question, then output only "No, I cannot answer". Otherwise, the comprehensive answer must contain citations to the source paragraphs, e.g., blablabla (paragraphs 1 and 2), blabla (paragraph 0). It should also start with "Yes" if the answer is positive, "No" if the answer is negative, or "N/A" if the answer is not available.*
> *Question: {question}*
> *Paragraphs: {contents}*

This template is designed for re-elaborating the output of an automated answer retrieval system like the one we employed. The prompt consists of guidelines on how to form a comprehensive answer based on the paragraphs provided for a specific question. It asks the system to generate an answer solely based on the information found in these paragraphs. It also requests that the source paragraphs be cited in the final answer for transparency. If no paragraph can answer the question, then the system is instructed to output 'No, I cannot answer'.

To further dissect the instructions:

- 'Output a comprehensive answer based only and exclusively on the information within the paragraphs below' means the algorithm should strictly use the provided text.
- 'if any can be used to answer' directs the algorithm to check the relevancy of the provided text snippets.
- 'If no paragraph can answer the question, then output only 'No, I cannot answer'' serves as a guideline for cases where the snippets do not have the required data.
- The directive that the answer 'must contain citations to the source paragraphs' mandates the algorithm to reference its sources, ensuring transparency and reliability.

The phrase 'extracted from the documentation to be assessed' is a key instruction aimed at clarifying the scope of where the answer should come from. This instruction is particularly necessary for cases where the question might have an implicit answer within the given documentation, but not an explicit one. In other words, it helps ChatGPT make an inference based on the information that is available, rather than stating 'I cannot answer' simply because an explicit answer is not given.

Once this filtration and aggregation process is complete, we apply the DoX metric to assess the explanatory depth of these refined answers.

The culmination of this process is the *explanatory relevance score*, which is calculated using the formula:

Explanatory Relevance Score = DoX Score×Max(Pertinence Score)

Here, 'Max(Pertinence Score)' is essentially the highest pertinence score among the retrieved correct answers for a checklist question.

---

[4]The version of GPT-4 used is 0613, a snapshot of GPT-4 from June 13th 2023.

The explanatory relevance score thus stands as an aggregate metric that captures both the depth of explanations and the specificity of content in terms of compliance with the regulations embodied in the checklist considered.

6 RESULTS

As suggested by the qualitative analysis presented in Section 4, platforms may exhibit varying degrees of compliance regarding ranking transparency. To verify this hypothesis, we went through a manual assessment of the six selected platforms' documentation, seeking to understand how well they comply with the level of detail and completeness of information required by the P2B Regulation. Then, we compared the results from the manual assessment with those obtained by using the automated assessment tools presented in Section 5.

For our manual evaluation, we employed two distinct methods. The first approach involved the expertise of three legal professionals who reviewed all relevant documentation and checklist inquiries. Conversely, our second method engaged over 130 participants. Given the associated costs, this method was applied to just two platforms and was narrowed to only four key questions.

The data mentioned in this paper and the code used for this experiment are available at [30].

6.1 Comprehensive Review by Legal Experts

Three Belgian senior legal experts each independently rated the documentation of the six platforms identified using the checklist from Section 4.2. Among these experts, two are academic researchers specializing in EU law on information and communication technologies, with respectively 7 and 5 years of experience in this field. The third, also well-versed in the same domain, with 6 years of experience, is a co-author of our study. To avoid biased results from this legal expert, as well as from the other two, three mitigation measures were taken. First, the set of questions of the checklist was elaborated objectively, before the evaluation of platforms' documentation. Second, the scores attributed to platforms' documentation during the assessments were based on an objective scale, described below. Third, legal experts compared and discussed their individual assessments after they were collected, to verify whether bias or other significant discrepancies appeared in the results of the co-author of the study, which was not the case.

These experts were instructed to answer the checklist exclusively based on the link sets identified in Section 4.1. Table 1 provides more details on the size of these link sets.

The experts were requested to use a scale from 1 to 5, defined as follows:

- **1**: The question is not answered at all.
- **2**: Indirectly or very poorly answered.
- **3**: Partly answered.
- **4**: Quite good, but not fully sufficient *vis à vis* the legal standard.
- **5**: Satisfactory – not necessarily perfect but close to the legal standard.

This scoring system allows for easier averaging across both experts and platforms, providing an aggregate measure of compliance.

During this first manual assessment, we rapidly found out that substantial differences existed between several platform operators. As illustrated in Table 2, Bing led with an average compliance score of 3.5. It was followed in descending order by Tripadvisor, Amazon, Google, Booking, and Yahoo.[5] Given that a score of '2' indicates a vague or indirect response, our findings resonate with the observations made by the European Commission about the quality of these explanations (Section 2).

Yahoo registered the smallest average variance in scores (0.12), indicating minimal disagreement among reviewers. Conversely, Bing and Google exhibited the most variance. As per Table 1, Yahoo and Bing/Google respectively had the fewest and most links under consideration, one might deduce that the volume of documentation influences expert ratings. This makes intuitive sense, as processing and retaining information from extensive texts can be challenging.

Table 2: Experts' Assessment: Average compliance scores given by the three experts for each platform, grouped by type and sorted by score. The scores are averaged across all the experts and all the questions for each platform.

Type	Platform	Average Compliance Score
Intermediary Service	TripAdvisor	2.627
	Amazon	2.569
	Booking	1.745
Search Engine	Bing	3.544
	Google	2.088
	Yahoo	1.228

Relying on only three legal experts, albeit realistic given the expense of specialists in the P2B Regulation, could skew interpretations. The small sample size of experts might not cover the breadth of possible scenarios or the diversity of people's opinions. Therefore, to mitigate this issue and corroborate the preliminary insights with more empirical data, we initiated a broader empirical study, focusing on evaluating software documentation from leading online intermediation services, namely Tripadvisor and Booking.

6.2 Large-scale Manual Assessment

Our larger scale empirical study involved 134 non-expert participants, sourced through Prolific.[6] Following the same methodology as the study conducted with experts (Section 6.1), participants were asked to evaluate the documentation identified for Tripadvisor and Booking presented in Section 4.1. They rated, on a 1-5 scale, how effectively the documentation answered questions related to the explanatory information mandated by the P2B regulations.

Guided by the European Commission guidelines (Section 3.1) stating that explanations about the ranking system should cater to the 'average' users' technical aptitude and needs for a given service (Section 3.1), participants were instructed to adopt the perspective of business users employing online intermediation services or online search engines.

---

[5]The low score for Yahoo is attributed to its reliance on Bing's documentation, as our assessment focused solely on Yahoo's own documents.
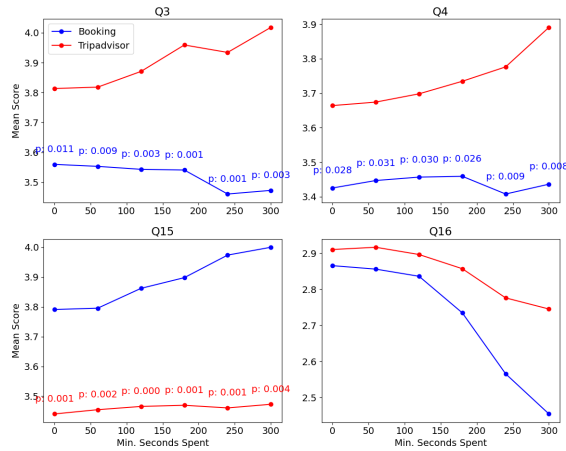[6]https://www.prolific.co

Figure 1: Average agreement rates for participants reviewing documentation on Tripadvisor and Booking, considering review durations from five to zero minutes. Includes statistically significant p-values for agreement rate comparisons between platforms.

To maintain the study's feasibility and ensure a median completion within approximately 10 minutes, we considered only four of the 17 questions presented in Section 4 and limited the analysis to the documentation of two intermediation services: Booking and Tripadvisor. These were chosen based on the brevity of their explanations (Booking is seven webpages long, Tripadvisor nine) and were representative of the best and worst explanations among the selected set of online intermediation services, as adjudged by the legal experts.

We curated the four questions, drawing from the list in Section 4, by considering the objective requirements of the regulation and its main goal, i.e., being able to understand how the ranking works and understanding how to improve/change the results of the ranking to improve outcomes.

Eventually, the selected questions, modified to fit a 1 (strongly disagree) to 5 (strongly agree) scale, were:

Q3 **Clarity of Ranking Mechanism:** The documentation clearly and sufficiently explains the mechanics and main parameters of the ranking system. Do you agree?

Q4 **Rationale for Ranking Parameters:** The documentation satisfactorily clarifies why certain parameters are critical in determining ranking instead of others. Do you agree?

Q15 **Improving Ranking:** The documentation adequately instructs users on how to improve the ranking of their goods or services. Do you agree?

Q16 **Altering Ranking Through Payments:** The documentation transparently discloses the paid options to improve ranking, and the effects thereof. Do you agree?

The study was designed to have evenly distributed male and female participants. Participants were only allowed from the UK[7]

---

[7]Since after Brexit, the UK has the discretion to retain, modify, or discard the provisions of the P2B Regulation.

or Ireland which are the only English-speaking countries where the P2B Regulation applies. We applied further pre-screening on the spoken language (participants had to be fluent in English) and the Prolific approval rate [13] which had to be above 75%.

Eventually, the study results aligned with our initial expectations, with observed differences between the perceived quality of explanations in the documentation of Tripadvisor and Booking. Furthermore, a few one-sided Mann-Whitney U tests also confirmed that these differences are statistically significant in most cases, as summarized in Figure 1:

- Tripadvisor surpasses Booking in detailing the ranking mechanism (Q3) and its rationale (Q4). The effect sizes for these observations are 0.42 and 0.44, respectively, suggesting a moderately strong relationship.
- Conversely, Tripadvisor lags behind Booking when it comes to guidance on enhancing rankings (Q15). This observation has a notably larger effect size of 0.60, indicating a stronger relationship.
- Regarding the transparency of paid ranking improvement options (Q16), the effect size is 0.49, suggesting a moderate relationship. Although no significant difference was measured, Tripadvisor still slightly outperforms Booking.

As shown in Figure 1, a deeper engagement with the documentation (achieved by spending more minutes) highlights the differences between Booking and Tripadvisor. That is because, intuitively, the longer one spends reading the documents, the more likely they are to notice nuances and differences.

However, only 55 of the 134 participants dedicated more than 5 minutes to reading the documentation. This points to the importance of providing brief and clear summaries at the start of documentation for quick information access. This would ensure that the most relevant information is efficiently communicated to non-expert users, within a reasonably limited timeframe.

The data in Figure 1 closely align with the experts' evaluations for each question. For instance, for Q3, experts ranked Tripadvisor at an average of 4.66, a score significantly higher than Booking's 2.33. In Q4, experts assigned Tripadvisor a rating of 3.66, outperforming Booking's 1.66. During Q15's assessment, Booking garnered a score of 4, marginally edging out Tripadvisor's 3.33. For Q16, experts gave Tripadvisor a 2.66, with Booking trailing at 1.

By cross-referencing the figures and experts' scores, there is a consistent pattern between the two, reinforcing the validity and reliability of the study findings.

Importantly, these findings were not only statistically significant but their effect sizes were medium to large, according to Cohen's conventional criteria. In particular, the effect sizes for Q3 and Q4 demonstrate a moderate practical advantage for Tripadvisor. While a large effect size for Q15 underlines the shortcomings of Tripadvisor when compared to Booking.

6.3 Automated Assessments

As shown so far, online platforms often struggle to produce clear documentation that complies with regulations like P2B. Evaluating and ensuring the quality of such documentation is challenging, largely due to the lack of standardized automated tools for assessing compliance and explainability across different platform providers.

Table 3: Comparative Statistical Data: Booking vs. Tripadvisor. Average scores are shown with their standard deviation. For each question, the best scores are shown in bold.

	Booking	Tripadvisor
Q3: Clarity of Ranking Mechanism	3.56 ± 0.95	**3.81 ± 0.86**
Q4: Rationale for Ranking Parameters	3.43 ± 1.04	**3.66 ± 0.96**
Q15: Improving Ranking	**3.79 ± 0.89**	3.44 ± 0.91
Q16: Altering Ranking Through Payments	2.87 ± 1.17	**2.91 ± 1.24**

To address this, we employed the assessment tools detailed in Section 5. Our aim was to determine if these tools' evaluations are consistent with expert opinions and the empirical findings discussed in Sections 6.1 and 6.2. We also compared their performance against baseline methods such as constant "Yes" responses and random[8] assessments.

The findings in Table 4 show that the DoX-based method and ChatGPT 3.5 match experts' evaluations 63.88% of the time. However, ChatGPT 4 is the least effective AI tool, performing even worse than random assessment. Since ChatGPT is a black-box, understanding why this occurs is impossible to us.

The scoring between the experts and ChatGPT was normalized as follows: any response scoring 3 or higher (indicating the question is "partially answered", as detailed in Section 6.1) was categorized as 'Yes', while scores below 3 were labeled 'No'. The DoX-based method already provides a straightforward 'Yes' or 'No' response.

Table 4: Agreements Rates: Percentage of outputs generated by different automated assessment strategies that align with that of the experts (Section 6.1). For example, for Bing, Chat-GPT 3.5 agreed with the majority of experts' answers 68.42% of the time. Scores in bold are the highest row-wise.

Platform	ChatGPT 3.5	ChatGPT 4	DoX-based	Random	Always Yes
Bing	68.42%	31.58%	68.42%	57.89%	**73.68%**
Booking	**76.47%**	29.41%	35.29%	35.29%	5.88%
Tripadvisor	58.82%	52.94%	**64.71%**	41.18%	58.82%
Amazon	47.06%	52.94%	**64.71%**	47.06%	**64.71%**
Google	36.84%	**57.89%**	52.63%	**57.89%**	31.58%
Yahoo	**94.74%**	**94.74%**	**94.74%**	42.11%	5.26%
All	**63.89%**	53.70%	**63.89%**	54.63%	39.81%

We also performed Mann-Whitney U tests and a rank biserial correlation analysis to analyze the alignment between the explanatory relevance, pertinence, and DoX scores produced by the second tool (Section 5.2) and the yes-or-no experts' answers.

As shown in Figure 2, all three scores show a significant p-value (less than .05), implying that the difference between the 'Yes' and 'No' groups in experts' answers is statistically significant for each score. The rank biserial correlation values range from -.456 to -.539, thus suggesting a moderate to strong negative correlation between the scores and the experts' answers. The explanatory relevance score correlates the best with the majority of answers, indicated by the highest absolute value of the rank biserial correlation (-.539).
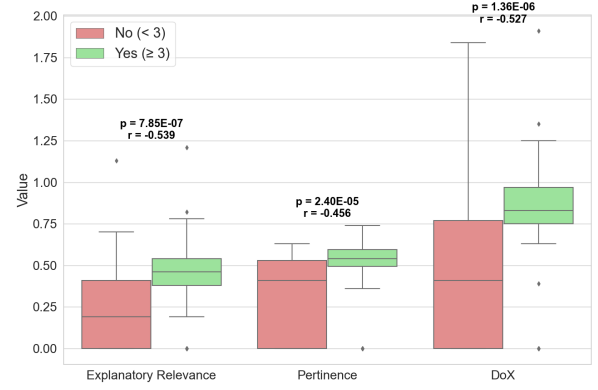
[8]We used 42 as a random seed.

Figure 2: Boxplot comparing 'Explanatory Relevance', 'Pertinence', and 'DoX' by experts' answers. Annotations indicate p-values and rank biserial correlation (r).

7 DISCUSSION & THREATS TO VALIDITY

The findings from Sections 6.1 and 6.2 highlight a statistically significant disparity in compliance scores across diverse platforms. This implies varying levels of compliance among platforms regarding ranking transparency. A possible reason for this inconsistency is the absence of standardized automated tools for gauging compliance and explainability.

Section 6.3 implies that automated tools could offer a solution. Notably, results indicate that ChatGPT 3.5 and the DoX-based approach, the best ones, align with the majority of experts 63.88% of the time. However, they are far from flawless, indicating there is considerable potential for improvement.

Delving deeper, the DoX-based approach seems to perform better than GPT-3.5 when evaluating longer documents. As illustrated in Table 1, for documentation from Google (52 docs, averaging 1679 words each), Bing (16 docs, averaging 964 words each), and Tripadvisor (10 pages, averaging 1653 words each), the DoX-based method achieves agreement rates of 52.63%, 68.42%, and 64.71%, respectively. In contrast, GPT-3.5 scores 36.84%, 68.42%, and 58.82%. These results can be attributed to the design of the DoX-based approach, which is optimized to handle long documents effectively, bypassing the input size constraints of ChatGPT.

We used ChatGPT without any fine-tuning. While the DoX-based approach incorporates GPT-4, its performance exceeds that of GPT-4 without answer retrieval. This hints at the instrumental role that prompt engineering might have in bolstering these tools' efficacy.

It is also crucial to understand the nuances of the DoX-based approach. Though this approach aligns with experts 63.88% of the time, this alignment merely considers ChatGPT's binary ('yes' or 'no') responses and omits the explanatory relevance scores. An in-depth examination of the alignment between the explanatory relevance scores from our DoX-based methodology and expert responses reveals their potential in pinpointing documentation sections that are non-compliant or inadequately articulated. Figure 2 illustrates that lower scores correlate with expert disapproval. Thus, the feedback depth from the DoX-based approach exceeds that of ChatGPT.

We acknowledge a potential limitation of this study: as discussed in Section 4.1, the study focused solely on textual content. However,

some platforms also feature video content or imagery that could provide additional insights. Yet, for intermediation services at least, it is compulsory to provide their explanations in written form (Section 4.1), hence our decision to focus on textual content. We cannot rule out that information not available in textual content could be embedded in videos or images.

Given the performance scores, the fully automated tools should not replace legal experts. These tools are effective at tracking flaws in documentation, but their best use is with legal professionals. Automated tools could be used to determine if existing documentation shows insufficient compliance, and if so, whether a legal expert is needed to enhance the problematic documentation.

Finally, Table 4 highlights cases in which ChatGPT 3.5 is the top performer, and cases in which the DoX-based method is more effective. Thus, for the best results, it seems reasonable to suggest combining both tools, then proceeding with manual validation.

8 SOCIETAL IMPLICATIONS

The qualitative assessment approaches to compliance measurement, and automated tools that we proposed in the previous sections have various societal implications, which we underline hereafter.

Our study provides concrete data and assessments of major market actors' documentation, in terms of compliance with a specific piece of EU law. It goes beyond existing literature by combining insights from legal experts, laypersons, and automated tools. This research can highlight ranking practices impacting individuals, businesses, and society. It can also assist platforms in understanding their level of compliance compared to competitors and addressing any issues. These results are beneficial for society.

Additionally, the paper highlights that assessing the quality of software documentation is a difficult, resource-consuming task, due to the lack of standardized automated tools to evaluate compliance. The paper also proposes a solution to this gap, in the form of automated assessment tools that are tested against human judgment. The societal benefits of this solution include enhanced means for regulators and public authorities to verify and force compliance with ranking transparency requirements. This, in turn, may lead to more balanced and fair markets for business and corporate website users, as well as more competition between platforms, as intended by the EU legislator. These benefits contribute to the United Nations Sustainable Development Goal 10.3 [25], aimed at reducing outcome inequalities and promoting equal opportunities.

Finally, our research explores approaches that can measure the quality of explanations provided about software and algorithmic processes *vis-à-vis* a legal standard, which is actually a crucial field of research. This is particularly due to the rapid development of *i)* machine learning techniques and tools, *ii)* methods to explain machine learning systems and their outputs [35], and *iii)* legal requirements imposing explanations when such systems are used [19]. In the context of our paper, it is clear that the ranking systems at stake include machine learning technology, which can be difficult to explain to laypersons. As lawmakers impose on online platforms to disclose information about these systems, it is particularly beneficial to society to be able to measure the quality of the explanations provided and determine when an explanation is insufficient or of poor quality. Besides, this reasoning and the

tools that we develop can be transposed to other contexts implying the compulsory provision of explanations in relation to machine learning tools or outputs, such as the forthcoming AI Act (see, for instance, Art. 13 of the proposed AI Act [11]).

9 CONCLUSIONS & FUTURE DIRECTIONS

The EU's P2B Regulation requires online platforms to clearly explain their ranking procedures. In this context, we studied how platforms like Amazon, Booking, Google, Bing, Tripadvisor, and Yahoo comply with this regulation.

With three legal experts and feedback from 130 users, we analyzed the documentation. Our findings showed varying compliance levels. One reason might be the lack of standardized assessment criteria or automated tools.

To address this, we introduced two automated assessment tools. First, given its popularity, we considered ChatGPT as an assessment method. Next, we proposed a hybrid method (designed to better handle long documentation) merging ChatGPT with answer retrieval techniques and a new metric, DoX, to measure documentation explainability.

When we compared the results from our automated tools to manual assessments, they agreed with human evaluations 63.88% of the time. This suggests these tools could help make the review process more efficient. Importantly, the method based on DoX was particularly good at assessing lengthy documents. Furthermore, its explanatory scores aligned well with expert opinions, implying that content with weak explanations tends to receive lower compliance ratings from professionals.

While these tools have limitations, their continuous assessment capability can be invaluable for regulators and the public. Combining them with expert judgment can promote transparent documentation, essential for a fair digital business space. This transparency aligns with goals like the UN's Sustainable Development Goal 10.3, fostering an informed and equitable online ecosystem.

Future improvements could refine these tools through better prompt engineering, tailoring models like ChatGPT for compliance tasks, and including multimedia analysis capabilities.

AUTHOR CONTRIBUTIONS

F. Sovrano and M. Lognoul: conceptualization, methodology, data curation, original draft preparation, investigation, validation, formal analysis. F. Sovrano: software, visualization. A. Bacchelli: methodology conceptualization, review, and editing.

REFERENCES

[1] Peter Achinstein. 2010. *Evidence, explanation, and realism: Essays in philosophy of science.* Oxford University Press.

[2] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in Chat-GPT: implications in scientific writing. *Cureus* 15, 2 (2023).

[3] Razvan Azamfirei, Sapna R. Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care* 27, 1 (March 2023), 120.

[4] Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29 (2021), 149–169.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 632–642. https://doi.org/10.18653/v1/d15-1075

[6] Erik Brynjolfsson and Brian Kahin. 2002. *Understanding the digital economy: data, tools, and research.* MIT press.

[7] Erik Brynjolfsson and JooHee Oh. 2012. The Attention Economy: Measuring the Value of Free Digital Services on the Internet. In *Proceedings of the International Conference on Information Systems, ICIS 2012, Orlando, Florida, USA, December 16-19, 2012*. Association for Information Systems. http://aisel.aisnet.org/icis2012/proceedings/EconomicsValue/9

[8] Christoph Busch. 2023. From Algorithmic Transparency to Algorithmic Choice: European Perspectives on Recommender Systems and Platform Regulation. In *Recommender Systems: Legal and Ethical Issues.* Springer International Publishing Cham, 31–54.

[9] European Commission. 2023. Report from the Commission on the first preliminary review of the P2B Regulation. https://digital-strategy.ec.europa.eu/en/library/report-commission-first-preliminary-review-p2b-regulation Accessed on: 2023-10-06.

[10] Content European Commission, Directorate-General for Communications Networks and Technology. 2020. Commission Notice Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council 2020/C 424/01. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020XC1208(01) Accessed on: 2023-10-06.

[11] Content European Commission, Directorate-General for Communications Networks and Technology. 2021. 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. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. Accessed on: 2023-10-06.

[12] European Parliament and Council. 2019. 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). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019R1150 PE/56/2019/REV/1, OJ L 186, 11.7.2019, p. 57–79.

[13] Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (2021), 1–20.

[14] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.

[15] Pablo A. Gajardo and Ariel I. La Paz. 2019. Measuring the strategic business and IT alignment in a digitally revolutionized economy. In *25th Americas Conference on Information Systems, AMCIS 2019, Cancún, Mexico, August 15-17, 2019.* Association for Information Systems. https://aisel.aisnet.org/amcis2019/org_transformation_is/org_transformation_is/19

[16] Golara Garousi, Vahid Garousi, Mahmood Moussavi, Günther Ruhe, and Brian Smith. 2013. Evaluating usage and quality of technical software documentation: an empirical study. In *17th International Conference on Evaluation and Assessment in Software Engineering, EASE '13, Porto de Galinhas, Brazil, April 14-16, 2013*, Fabio Q. B. da Silva, Natalia Juristo Juzgado, and Guilherme Horta Travassos (Eds.). ACM, 24–35. https://doi.org/10.1145/2460999.2461003

[17] Golara Garousi, Vahid Garousi-Yusifoglu, Günther Ruhe, Junji Zhi, Mahmood Moussavi, and Brian Smith. 2015. Usage and usefulness of technical software documentation: An industrial case study. *Inf. Softw. Technol.* 57 (2015), 664–682. https://doi.org/10.1016/j.infsof.2014.08.003

[18] Philipp Hacker, Johann Cordes, and Janina Rochon. 2022. Regulating Gatekeeper AI and Data: Transparency, Access, and Fairness under the DMA, the GDPR, and beyond. *arXiv preprint arXiv:2212.04997* (2022).

[19] Philipp Hacker and Jan-Hendrik Passoth. 2020. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers.* Springer, 343–373.

[20] Leah Henrickson and Albert Meroño-Peñuela. 2023. Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & SOCIETY* (Sept. 2023). https://doi.org/10.1007/s00146-023-01752-8

[21] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[22] Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for Explainable Automated Student Answer Assessment. *CoRR* abs/2305.12962 (2023). https://doi.org/10.48550/arXiv.2305.12962 arXiv:2305.12962

[23] Vesna Lukovic. 2021. INFORMATION ASYMMETRIES IN ALGORITHMS AT DIGITAL PLATFORMS: MOTIVATIONS TO PARTICIPATE AND EU REGULATORY APPROACH. *EMAN 2021–Economics & Management: How to Cope with Disrupted Times* (2021), 167.

[24] Catarina Dutilh Novaes and Erich H. Reck. 2017. Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synth.* 194, 1 (2017), 195–215. https://doi.org/10.1007/s11229-015-0816-z

[25] Department of Economic and Social Affairs. 2023. Goal 10: Reduce inequality within and among countries. https://sdgs.un.org/goals/goal10 Accessed on: 2023-10-06.

[26] Observatory on the Online Platform Economy. 2021. Study on "Support to the Observatory for the Online Platform Economy". https://platformobservatory.eu/app/uploads/2021/01/P2B-Regulation-monitoring-analysis-January-2021.pdf. Accessed on: 2023-10-06.

[27] Thalia Prastitou-Merdi. 2021. The Notion of "Online Intermediation Services" Found in the New EU Platform Regulation: Who Is Caught After All? In *EU Internet Law in the Digital Single Market.* Springer, 543–559.

[28] Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can ChatGPT Assess Human Personalities? A General Evaluation Framework. *CoRR* abs/2303.01248 (2023). https://doi.org/10.48550/arXiv.2303.01248 arXiv:2303.01248

[29] Shaping Europe's digital future. 2023. EU Observatory on the Online Platform Economy. https://digital-strategy.ec.europa.eu/en/policies/eu-observatory-online-platform-economy. Accessed on: 2023-10-07.

[30] Francesco Sovrano. 2023. Replication Package for "Automating Regulatory Compliance: An Empirical Study on Ranking Transparency in the Software Documentation of EU Online Platforms". https://github.com/Francesco-Sovrano/Automating-Regulatory-Compliance-An-Empirical-Study-on-Ranking-Transparency-of-EU-Online-Platforms. https://doi.org/10.5281/zenodo.10478546

[31] Francesco Sovrano, Monica Palmirani, and Fabio Vitali. 2020. Legal Knowledge Extraction for Knowledge Graph Based Question-Answering. In *Legal Knowledge and Information Systems - JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020 (Frontiers in Artificial Intelligence and Applications, Vol. 334)*, Serena Villata, Jakub Harasta, and Petr Kremen (Eds.). IOS Press, 143–153. https://doi.org/10.3233/FAIA200858

[32] Francesco Sovrano and Fabio Vitali. 2022. Explanatory artificial intelligence (YAI): human-centered explanations of explainable AI and complex data. *Data Mining and Knowledge Discovery* (2022). https://doi.org/10.1007/s10618-022-00872-x

[33] Francesco Sovrano and Fabio Vitali. 2022. How to Quantify the Degree of Explainability: Experiments and Practical Implications. In *31th IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padova, July 18-23, 2022.* IEEE, 1–9.

[34] Francesco Sovrano and Fabio Vitali. 2023. An objective metric for Explainable AI: How and why to estimate the degree of explainability. *Knowledge-Based Systems* 278 (2023), 110866. https://doi.org/10.1016/j.knosys.2023.110866

[35] Timo Speith. 2022. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 2239–2250.