December 2024

Automating Fundamental Right Impact Assessment: an Open Experiment

Xinyue ZHANG [a], Vanja SKORIC [a,1] and Giovanni SILENO [a]
[a] *University of Amsterdam, Amsterdam, the Netherlands*

Abstract. With the adoption of the AI Act, fundamental rights impact assessment (FRIA) processes become highly relevant for both public and private institutions; yet such processes can be challenging, especially for small- to medium-sized organizations. One recent research that piloted a partial automation of FRIA is Anticipating Harms of AI (AHA!), relying on the use of a large language model and crowd-sourcing; unfortunately, the paper provides limited insights upon its internal working. Therefore, this work presents AFRIA, a processing pipeline that performs specific aspects of FRIA, conceived with AHA! as inspiration. In order to assess to what extent AFRIA is a successful reconstruction of AHA!, we analyzed the percentage of meaningful harms that AFRIA generates and the distribution of harm categories, and compared it to AHA!'s results, finding a satisfactory convergence. Beyond inspiration from AHA!, we also looked into the requirements of the AI Act and scholarly critique to make AFRIA more meaningful for identifying impacts on fundamental rights, targeting categories of human rights impacts, potential harm mitigation measures, and the severity and likelihood of the harms. The results show opportunities, but also limitations in what type of support this technology can bring.

Keywords. AI impact assessment, fundamental rights impact assessment, large language models, automated processing, human-computer interaction

1. Introduction

In March 2024, the European Union (EU) adopted the Artificial Intelligence Act (AIA) to regulate product safety in light of the increasing utilization of Artificial Intelligence (AI). The AIA's introduction of fundamental rights impact assessment (FRIA) brings new requirements, including that deployers need to share FRIA results with the oversight body. While these requirements could increase citizens' safety, FRIA processes also bring potential challenges. To ensure inclusivity and to gain a variety of perspectives when conducting an impact assessment, the input of relevant external stakeholders and diverse experts is crucial [1,2]. Manual impact assessment can therefore be time- and cost- exhaustive.[2] This could pose potential challenges for both public and private institutions which may not have sufficient resources to conduct a proper impact assessment as required by the AIA.

[1]Corresponding Author: Vanja Skoric, `v.skoric@uva.nl`.

[2]Take for instance the *Impact Assessment Mensenrechten en Algoritmes* (IAMA) of the Netherlands [3]. This developing FRIA tool uses a questionnaire-based approach that divides the impact assessment into *why* ("why do we need AI?"), *what* ("what AI will we use?"), and *how* ("how will this AI be used?"). It requires in-depth research and fieldwork, as well as the input of experts (e.g. legal and human relations).

Therefore, it is relevant to look into resource-saving methods to conduct FRIA, at least in the explorative phases. Yet, the field of AI-powered FRIA is new, and only little research exists on the subject. A notable example is the recent project by Microsoft and Harvard called *Anticipating Harms of AI* (AHA!), which uses large language models (LLMs) to produce a generative framework to help practitioners anticipate the harms of AI usage [4]. However, the technical details of AHA! are not accessible.[3] Therefore, the present work presents a reproducible AI-powered FRIA tool (AFRIA), taking AHA!'s generative framework as inspiration. Furthermore, we attempt to extend AFRIA beyond AHA!, by generating other factors that might be relevant for FRIA: impacted fundamental rights, mitigation methods, and the severity and likelihood of the harms, following the debate in the legal scholarly literature on the subject.

Our research questions are: (RQ1) *Can AFRIA generate meaningful examples of harm?* (RQ2) *Do the categories of harms differ significantly depending on the scenario?* (RQ3) *Do the categories of harms differ significantly depending on the dimension of problematic AI behavior?* Performances on these tasks will be compared to those presented by AHA!. Additionally, we will investigate the following questions: (RQ4) *Can AFRIA generate meaningful categories of fundamental rights? For each harm, can it generate mitigation measures and estimations of severity and likelihood?*

This paper is structured as follows. First, we will provide background information: an outline of the criteria for FRIA, of large language models (LLMs), and details of the AHA! model. Afterward, we will discuss our methods for creating AFRIA and for analyzing its generations. We will then analyze the results to answer the research questions and conclude with the limitations and future recommendations.

2. Background

2.1. Fundamental Rights Impact Assessment (FRIA)

The AIA introduces stricter requirements for FRIA (see Article 27(1)), making it obligatory before deploying high-risk AI products. It requires a FRIA to assess AI's risks to fundamental rights by mapping out several factors: the deployer's processes, the frequency of the AI usage, the categories of people that may be affected, the potential harms, intervention methods and risk mitigation measures.[4] With the AIA, watchdogs and protection authorities have more tools to hold AI deployers accountable. Therefore, conducting FRIA has become more relevant for deployers, and, consequently, for the providers. Yet, Mantelero [6] convincingly observes that the AIA does not sufficiently cover all aspects a FRIA should cover. He identifies two types of impact assessments: *awareness-raising* and *risk-based*. The former is in line with Article 27 of the AIA, and is based on contextualising the AI and its usage, for example, in which scenario the AI system is utilized and which stakeholders are affected. One example of the awareness-raising methodology is IAMA [3]. Mantelero, however, critizes it for failing to adequately assess potentially affected human rights, and placing instead its focus on the needs of the user and the product's features. A risk-based methodology, on the other hand, concerns the AI's level of

[3] AHA!'s code is not publicly available, and the paper does not provide details of the code, nor of the prompts.

[4] Previous EU regulations, like the GDPR, were less specific. The GDPR, for example, is criticized for its lack of specifications to keep data controllers accountable for their own efforts to self-regulate [5].

risk on impacted fundamental rights and its development in relation to mitigation measures and changes in technology, society, and the scenario in which it is used [6,7]. This methodology is not included in the AIA, but Mantelero stresses its importance because it underlines the risk's dependency on dynamic contextual factors. Two dimensions are key to risk assessment: the *severity* and *likelihood* of the risk. Ideally, FRIA should assess these two dimensions for every impacted right.

2.2. Large Language Models (LLMs)

LLMs excel in different linguistic tasks across domains [8,9]. They are trained on large data sets and have high model complexity. Because their interpretability is poor, and their workings remain hidden behind a black box, it is important to evaluate their performance, trustworthiness, and robustness, and research is ongoing on these dimensions. In general, LLMs do very well on text generation tasks by producing fluent and precise linguistic expressions [8]. Yet, LLMs perform poorly on abstract reasoning and are prone to confusion in complex contexts [8]. LLMs are also prone to fabricating information, a phenomenon commonly referred to as "hallucination" [10]. Some authors argue that LLMs cannot reason at all because they are purely information retrieval machines [11]. LLMs perform poorly in robustness, as the same inputs lead to different outputs, especially when the inputs use different expressions and grammar [8]. Due to their training data and retrieval functions, LLMs can enforce societal biases [12]. Bias tends to grow in size as the model size increases, posing challenges for the big five families of LLMs: GPT, OPT, BLOOM, LLaMa, and BlenderBot [13]. However, research also commends black-boxed AI's potential to minimise bias. The construction of models can be tested and scrutinized, but we cannot do the same with human biases [14].

2.3. Anticipating Harms of AI (AHA!)

In the literature, a few LLMs-based methods have been proposed to anticipate AI harms by supporting model auditing [15,16], or assessing the risk or impact in scenarios [17,18,19]. AHA! [4], however, is the only model to date that piloted an automated solution for impact assessment procedures by generating harms for different stakeholders using an LLM (GPT-3). The pipeline of AHA! functions as follows. First, the user prompts the model with the *scenario* in which the AI is used, specific *stakeholders* that the user identified, and the problematic AI behaviours (hereafter referred to as *harm dimensions*) the user is interested in. The authors have tested the following harm dimensions: *False positives (FP)*: the system predicts a positive outcome when it should not; *False negatives (FN)*: the system predicts a negative outcome when it should not; *One-time mistake*: the system makes a one-time error; *Accumulated mistake*: the system makes an error repeatedly or systematically; *Egregious severity*: the system makes a severe error; *Unspecified severity*: the system makes an error of unspecified severity; *Conditioned on a specified harm*: the vignette is conditioned on a specific harm (e.g. financial harm for the hiring scenario); *Unspecified harm*: the vignette is not conditioned on a specific harm.

Based on the scenario, relevant stakeholders, and harm dimensions provided by the user, GPT-3 is used to generate a larger list of potential stakeholders. These stakeholders are entered into the rows of an "ethical" matrix, with the problematic AI behaviour as columns. Then, GPT-3 is used to generate a *vignette* for every cell of the matrix. A vi-

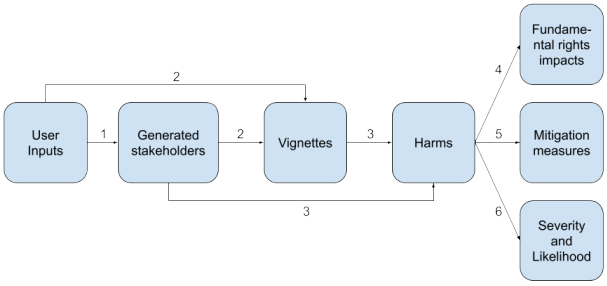

Figure 1. AFRIA's pipeline, simplified.

gnette is a fictive sub-scenario the stakeholder may experience. An example of a vignette in the hiring scenario is as follows: "*... Despite having relevant experience and skills for the position, the AI hiring system mistakenly categorizes your resume as not suitable for the job.*" After generating the vignette, AHA! generates the harm that the stakeholder could experience for that vignette. It does so in two ways: by crowdsourcing and GPT-3 generation. Lastly, the different harms are categorised manually by researchers.[5]

3. Methods

3.1. AFRIA Framework Design

The AFRIA pipeline we propose has been conceived building upon the available information around AHA!, extended with additional categories. It realizes a generative framework that could potentially help practitioners identify related stakeholders, anticipated harms, fundamental rights impacts, potential mitigation methods, and the severity and likelihood of harms.[6] Like AHA!, the user is asked to provide three inputs at the start: the *scenario*, relevant *stakeholders*, and *harm dimensions*. For example, a *scenario* is a tech company that wants to use AI in the hiring process to assess whether applicants are a good fit for the job. Furthermore, the user gets the option to make AFRIA generate a dimension of specific harms (e.g. if the user only wants to generate financial harms). Clearly, for better generations, the user needs to provide inputs as refined as possible. This is also because the prompts AFRIA rely upon need to be generalisable across different scenarios and harm dimensions. See Table 1 for the specific prompts used.

After the user has provided the inputs, the LLM's chat completion will generate another list of stakeholders (step 1 in Fig. 1), based on the scenario and the initial list. AFRIA is prompted to take the input of stakeholders as an inspiration, meaning that it does not have to include all the identified stakeholders in its generated list. This way,

[5]Interviews with 9 responsible AI practitioners/academics resulted in different views on AHA!'s utility [4]: some praised its ability to cover more potential harms and stakeholders, and saw AHA's potential in convincing decision-makers about AI's dangers with long lists of harms. Critics, however, fear that practitioners will become overly reliant on AHA!. Some practitioners wanted more support in mitigating harms and assessing their severity, which aligns with Mantelero's emphasis on the importance of risk-based methodologies [6].

[6]Our experiments with AFRIA have been implemented with simple Python scripts that interact with the user and interrogates the LLM via an API. The code, relying on OpenAI's API (to compare the performance of our pipeline with that reported with AHA!), is publicly available (see `https://github.com/XCINDYZ/AFRIA`). Users can easily experiment with other LLMs by changing the API.

Task	Prompt
Generated stakeholders	"Come up with a list of potential direct stakeholders and a list of potential indirect stakeholders in a scenario where AI is used: {scenario}. Take inspiration (and include relevant stakeholders) from this list: {specific_stakeholders}. Direct stakeholders are considered those who directly interact with or are immediately affected by an AI system, while indirect stakeholders may be people associated with direct stakeholders or larger community groups."
Vignette	"Narrate how {stakeholder} in the scenario ({scenario}) may experience {behaviour}. Formulate your answer in second-person perspective: 'Imagine you are a [stakeholder], you may experience [harm] because..."
Vignette of specified harms	"Narrate how stakeholder in the scenario (scenario) may experience this problematic AI behaviour: ({behaviour}). Formulate your answer in second-person perspective: 'Imagine you are a [stakeholder], ..."'
Harm	"Summarise the vignette ({vignette}), and specify the harm the stakeholder faces due to the problematic behaviour AI behaviour. Formulate your answer in second-person perspective: 'Imagine you are a [stakeholder], ..."'
Fundamental rights impact	"What fundamental rights (enshrined in the universal declaration of human rights) of the stakeholder '{stakeholder}' are affected by this harm: {harm}. Refrain from mentioning rights and freedoms of other stakeholders, but only focus on the rights and freedoms of {stakeholder}."
Mitigation measures	"Given the harm ({harm}) faced by {stakeholder}, propose mitigation measures."
Severity	"Assess the severity of the harm: {harm}."
Severity level	"For the harm '{harm}', assign the level of the severity on a scale of low, medium, high, very high: {severity}"
Likelihood level	"For the harm '{harm}', assign the likelihood of the harm happening on a scale of low, medium, high, very high."
Likelihood confidence	"How confident are you about the assessment of the likelihood for the harm: {likelihood}"

Table 1. Prompts used in AFRIA to interrogate the LLM

AFRIA also considers stakeholders the user did not think of and filters out irrelevant stakeholders the user identified. AFRIA will then fill a "harm matrix", with as rows the generated stakeholders, and as columns the harm dimensions that the user had provided as input (e.g. "false positive"). Now, AFRIA will start populating each harm matrix cell (per stakeholder per harm dimension) with relevant information. First, per cell, chat completion is prompted to generate a vignette (step 2 in Fig. 1). This is a fictive scenario the stakeholder may experience, which can be understood as a sub-scenario. For example, in the hiring scenario, that a hiring applicant does not get the job at the tech company due to a FN decision.

For every vignette, AFRIA will summarize it for the user and be prompted to specify the harm the stakeholder could face in the vignette (see step 3 in Figure 1). These harms can be multifaceted, ranging from financial harms to societal ones. One harm generation can contain multiple harms of different types. This generation, with a summarized vignette, is put into the matrix cells.[7]

[7] Unlike AHA!, which combines this process with generated harms by crowds and LLM, AFRIA only generates harms through an LLM.

During the input process, the user could also specify if they want to generate additional information concerning fundamental rights impacts, harm mitigation measures, and severity and likelihood. In this case, AFRIA will generate these additional information on the basis of the harm matrix, creating additional fundamental rights, mitigation, and likelihood and severity matrices. Note that in the prompt we specified that the model should refrain from generating the rights impacts of other stakeholders, as we discovered it improved the fundamental rights impacts output.

3.2. Experimental Methods

In this work, we used OpenAI's API to run AFRIA. This choice was made in order to produce results that are as similar as possible to AHA!, which used GPT as the LLM model [4]. More precisely, the only specification was that `GPT-3 davinci` model was used to generate the harms. However, `GPT-3 davinci` model is deprecated, and no longer available. At the time of this research, OpenAI suggests `GPT-3.5-turbo` as the substitute for `GPT-3 davinci`.[8] To compare our results with AHA!'s results, we used the same five scenarios and their corresponding stakeholders of the original paper ([4], p. 19). We also used the same harm dimensions (see section 2.3). First, regular false positive (FP) or false negative (FN). The rest of the scenarios are specified variations of FP and FN: one-time/accumulated FP/FN, egregious/unspecified FP/FN, specified/unspecified FP/FN harm. For this experiment, we used the following specific harms: in the hiring and loan application scenario, the specified harm is financial strain. For communication compliance and content moderation, the specified harm is emotional strain. For disease diagnosis, it is health concerns.

3.3. Coding and analyzing harms

Because LLMs generate (slightly) different results every run, it is difficult to decide upon a number of executions adequate to generalise the results. Therefore, we kept the number of harm generations to produce results similar to AHA! (4113 harms). We matched this by generating 2189 harm matrix cells across all five scenarios, which eventually resulted in 4580 harms spread over the categories.

To answer to RQ1, we encode harms in a similar way as done by AHA! ([4], p. 8), yet, as we found the description not always clear, we rephrased it differently. A harm cell is categorized as *meaningful* if all these three conditions are satisfied: (i) the connection between action and consequences is plausible or logical; (ii) the consequences have actual harmful effects; (iii) the harm is affecting the target stakeholder. Vice-versa, *non-meaningful* harms are ones in which any of the three conditions is not satisfied. Non-meaningful can be distinguished as *non-sensical* (when the connection action-consequences is implausible/illogical), or *sensical* otherwise.[9] When a cell contains both

[8] We also briefly experimented with different LLMs via Ollama, such as `Mistral` and `llama-2`. However, those tended to converse more with the user, therefore resulting in repeating questions or giving more unsolicited mitigation advice, being sub-optimal for the current pipeline.

[9] Example of non-sensical harm: *"Imagine you are the employee receiving the [toxic] emails. The harm you face [...] includes frustration and anxiety when the system fails to detect your use of toxic language..."*. Sensical (non-meaningful): *"Imagine you are a consultant [...] You discover a pattern of false positives where the AI incorrectly predicts certain applicants as good fits for the job when they lack the necessary qualifications. This issue can lead to hiring candidates who are not truly suitable, causing inefficiencies, performance issues..."*

meaningful and non-meaningful harms, we coded the cell into both categories. It is also possible that the harm is non-meaningful because the vignette is non-meaningful. Therefore, we also recorded the number of non-meaningful vignettes. When the cell is non-meaningful due to a non-meaningful vignette, the cell will not be considered in the calculation of meaningful harms.

For RQ2 and RQ3, we coded the matrix cells into the harm categories developed by AHA! ([4], pp. 21–22). A matrix cell can contain multiple harm categories. For example, the harm *"...the applicant experiences distress and missed opportunities"* contains two categories of harm: well-being and allocational harm. In order to analyze the distribution of harms, the harms are coded into both categories. If a cell has multiple harms from the same category, we count it only once. When the cell contains un-prompted mitigation methods, we ignore them in our coding. Through this process, the 2189 matrix cells are split into 4580 harms, which is similar to the number of harms AHA! based their results on. To minimize coding errors, we reviewed our categorization twice. Because manual coding was time-consuming, we also attempted automated categorization with and without few-shots prompting. Then, to answer RQ2, we created a heatmap of the distribution of harm categories across scenarios. To assess whether the distribution of harm categories is significant, we conducted a chi-square test with a post-hoc chi-square pairwise comparison using the Holm-Bonferroni method [20]. To answer RQ3, we conducted a chi-square test for the distribution of harms across harm dimensions in 3 ways: comparing the distribution of FP and FN dimensions, comparing the distribution of accumulated/one-time and egregious/not specified harms, analyzing if specifying the harm generates significantly more harms of that kind.

3.4. Coding and analysing fundamental rights

AFRIA also generates fundamental rights impacts, mitigation measures, and the severity and likelihood of harms. This work prioritized analyzing the fundamental rights impacts, and to what extent it can generate meaningful results. The initial idea was to generate around 4113 fundamental rights impacts to match AHA!'s harm generation. However, during experimentation, we found that many of the fundamental rights impacts were non-meaningful. Nonetheless, we coded the generations to better understand the amount of non-meaningful results and the distribution of rights. To quantitatively analyze this, we generated 168 impacted fundamental rights. We will not conduct a significance test of the distribution because that would be meaningless due to the small number of outputs.

First, we categorized the fundamental rights matrix cells into *meaningful* and *non-meaningful* categories. A fundamental right impact is meaningful when it is an actually impacted right to the stakeholder, that exists in the scenario in question (see properties i, ii, iii for the meaningful harm categorization). We also counted as wrong cases in which the model labels a right incorrectly.[10] This is done in consideration of the potential usage of AFRIA in the absence of legal experts to correct inaccurate labels. Non-meaningful rights are categorised into nonsensical rights and sensical (but irrelevant). Nonsensical rights are those that are not clear or do not make sense in the context.[11] When a cell con-

[10] For example, a description of the right to health which the model referred to as the right to life.

[11] An example of a non-meaningful right is as follows: "*Right to non-discrimination of the manager: [...] there is a risk of unintentional discrimination against qualified candidates, violating the hiring manager's responsibility to uphold equal opportunities.*" While the generation of the LLM makes sense, it is non-meaningful because its description does not indicate that the right to non-discrimination of the manager is violated.

	Communication compliance	Content moderation	Disease diagnosis	Hiring	Loan application
Quality-of-service	**<0.001**	0.1936 (n.s.)	**<0.001**	0.0027 (n.s.)	0.1336 (n.s.)
Representational	0.00373 (n.s.)	**<0.0001**	**<0.0001**	0.0719 (n.s.)	1.0 (n.s.)
Well-being	0.6171 (n.s.)	**<0.001**	**<0.0001**	0.0357 (n.s.)	**<0.0001**
Legal-reputational	0.6892 (n.s.)	0.1336 (n.s.)	**<0.0001**	**<0.0001**	0.1615 (n.s.)
Social-societal	**<0.0001**	0.4237 (n.s.)	**<0.0001**	**<0.0001**	**<0.0001**
Loss of rights	0.2391 (n.s.)	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
Allocational	**<0.0001**	**<0.0001**	0.7642 (n.s.)	**<0.0001**	**<0.0001**
Other	0.6892 (n.s.)	0.7642 (n.s.)	0.6892 (n.s.)	0.4839 (n.s.)	0.7642 (n.s.)
Non-meaningful	0.1096 (n.s.)	**<0.001**	0.0093 (n.s.)	0.6892 (n.s.)	**<0.0001**

Table 2. χ^2 post-hoc analysis with Holm-Bonferroni corrections of the harm distribution.

tains both meaningful and non-meaningful fundamental rights impacts, the cell is categorised in both categories. We then categorised the different forms of fundamental rights impacts, and the categorization was revised by a legal expert (one of the co-authors). This was challenging because we did not have an existing established taxonomy to consult. We created our own taxonomy by open-coding the generated rights impacts into categories. To limit the variety of rights to make manual coding easier, we prompted the model to generate rights based on the fundamental rights in the *Universal Declaration of Human Rights* (UDHR). However, despite the explicit prompt, AFRIA still generated other rights. Since we had to capture a wide variety of rights, we open-coded the generations into wide categories before grouping them into taxonomies. This is the same approach used to create the harm taxonomy in [4]. We also tried to automate the coding process by prompting the LLM. To answer RQ4, we calculated the percentage of non-meaningful rights impacts and displayed their distribution with a heatmap.

4. Results

Our attempts at automating the categorization processes (for harms and fundamental rights) via the LLM were not successful.[12] We had to rely then on manual coding.

RQ1: Can AFRIA generate meaningful examples of harms? Of the 2189 matrix cells, 90.9% were generated from meaningful vignettes. Of all the cells generated from meaningful vignettes, 93.7% contained meaningful harms. Of the 6.3% non-meaningful harms, 95.1% were nonsensical. In comparison, AHA! found 93% meaningful harms (elicited aggregating crowds and GPT-3 outputs). Of the 7% non-meaningful harms, 86.4% were nonsensical. Thus, AFRIA's percentage of meaningful harms is comparable (0.7% higher) to AHA!. The proportion of nonsensical harms is also similar.

RQ2: Do the categories of harms differ significantly depending on the scenario? Since we generated harms across different scenarios, we hypothesised that different scenarios would generate different harm categories. To test this hypothesis, we first created a heat map of the distribution of harm categories. The map shows that for communication compliance and content moderation, AFRIA generates more well-being and legal-and-reputational harms; for disease diagnosis, more well-being and legal-and-reputational,

[12] For the harms, we hypothesize that is because the harms are too multifaceted and context-dependent. We had higher expectations for the fundamental rights impacts because AFRIA's rights impacts generation would often mention the name of the rights in question, and so it could have been easier for the LLM to categorise fundamental rights. However, we were unsuccessful, because AFRIA generated a wide variety of rights beyond those of the UDHR and often labelled descriptions wrongly.

	Communication compliance	Content moderation	Disease diagnosis	Hiring	Loan application
Quality-of-service	0.76418 (n.s.)	0.13361 (n.s.)	**<0.001**	0.42371 (n.s.)	0.27133 (n.s.)
Representational	0.42371 (n.s.)	**<0.0001**	0.03573 (n.s.)	0.27133 (n.s.)	0.48393 (n.s.)
Well-being	0.01242 (n.s.)	**<0.001**	0.27133 (n.s.)	**<0.0001**	**<0.0001**
Legal-reputational	1.0 (n.s.)	0.23014 (n.s.)	0.07186 (n.s.)	0.07186 (n.s.)	0.1615 (n.s.)
Social-societal	**<0.0001**	0.31731 (n.s.)	**<0.001**	0.07186 (n.s.)	0.1936 (n.s.)
Loss of rights	0.36812 (n.s.)	**<0.0001**	0.08913 (n.s.)	0.08913 (n.s.)	0.13361 (n.s.)
Allocational	**<0.0001**	**<0.0001**	0.08913 (n.s.)	**<0.0001**	**<0.0001**
Other	0.54851 (n.s.)	0.61708 (n.s.)	0.61708 (n.s.)	0.61708 (n.s.)	0.0164 (n.s.)
Non-meaningful	0.05743 (n.s.)	0.31731 (n.s.)	0.84148 (n.s.)	0.0455 (n.s.)	0.27133 (n.s.)

Table 3. post-hoc χ^2 pairwise analysis of the distribution of specified harms across scenarios

and quality-of-service harms; for hiring and loan application, more allocational harms. This is rather similar to the results of AHA! We want then to test whether differences were significant. To assess this, we ran a chi-square post-hoc test with Holm-Bonferroni correction. Some of the differences are significant (see Table 2). Notably, hiring and loan application scenarios surface significantly more allocational harms. Likewise, disease diagnosis surfaces significantly more well-being and legal-and-reputational harms. However, while communication compliance and content moderation surface more legal-and-reputational harms, this distribution is not significant. Instead, communication compliance surfaces more social and societal harm which is significant. This could be because the AI is used in a work environment, and is therefore able to impact the relationship between employees and the work environment. Content moderation surfaces more well-being harm which is significant. This could be because the AI is used to remove harmful social media posts. If the AI produces harmful behaviour, it could be emotionally distressing for social media users.

RQ3: Do the categories of harms differ significantly depending on the dimension of problematic AI behavior? To assess if the harm distribution varied significantly across different dimensions, we conducted chi-square analyses across scenarios Like AHA!, we found no significant differences for one-time vs. accumulated or egregious vs. unspecified harms. However, while AHA! identified significant differences in harm categories for false positive vs. false negative dimensions in most cases, AFRIA only showed this significance in the content moderation scenario.[13] To analyze whether specifying harms surface more harms of that category, we generated a heat map consisting of only the harms surfaced by the specified harm dimension. Communication compliance and content moderation, which were prompted with emotional distress, surfaced more well-being harms. Disease diagnosis, which was prompted with health concerns, surfaced more well-being harms as well. Hiring and loan application, which were prompted with financial concerns, surfaced more allocational harms. These observations coincide with AHA!'s findings, except for the disease diagnosis scenario (where AHA! found many quality of service harms). To see if these distributions were significant, we ran a post-hoc chi-square pairwise comparison again of the harm distribution (see Table 3). The observation above is significant for content moderation, hiring, and loan application. Quality-of-service is also significantly dependent on the disease diagnosis scenario. Well-being is not significantly dependent on communication compliance and disease diagnosis.

[13]For example, false positives (an AI mistakenly flagged an email) led to more quality-of-service harms, while false negatives (the AI failed to detect a toxic email) resulted in more social and societal harms.

RQ4: Can AFRIA generate meaningful examples of fundamental rights? Of the 168 rights cells, 8.9% are generated from non-meaningful vignettes. Of the cells generated from meaningful vignettes, 58.2% are meaningful. This result is disappointing, considering that AFRIA generated 93.7% meaningful harms. Of all the non-meaningful impacted fundamental rights, 12.1% are sensical and 87.8% are nonsensical. The heatmap (here not reported for space reasons) shows that communication compliance and content moderation scenarios generate more expression-related rights, like freedom of expression. Disease diagnosis generates more well-being-related rights, such as the right to health. Hiring and loan applications generate more work-related rights, which are often related to benefit distribution. This shows AFRIA's ability to generate relevant fundamental rights impacts depending on the scenario.

Report of the Mitigation Methods and the Severity and Likelihood AFRIA also generated the mitigation measures and the severity and likelihood of the harms. At the moment, we conducted only a preliminary qualitative analysis on these outputs: the generated mitigation measures overall made sense; severity and likelihood were instead unsuccessful, generating high severity and likelihood of harms for almost all cases.

5. Discussion

This work offered insights into the potential and the challenges of partly automating FRIA, providing an LLM-based tool (AFRIA) to generate the harms that stakeholders face given a scenario. We draw our inspiration from AHA!'s pipeline, but we also extended it, targeting impacted fundamental rights, mitigation measures, and the severity and likelihood of the harms. Overall, using `GPT-3.5-turbo`, AFRIA confirmed performances observed with AHA! (which includes also crowd-sourcing). For the five scenarios under study, it was able to generate meaningful harms (93.7%), which significantly differ across scenarios. Yet, with AFRIA, we found less significant differences in harm across harm dimensions (FP vs FN), except for the content moderation domain. This could be due to our prompting or due to inherent limitations of `GPT-3.5-turbo`. Besides, AFRIA's performance in generating meaningful fundamental rights (i.e. actual and relevant to the context) was much poorer (58.2%). Performing a preliminary analysis on additional components, we found that the suggested mitigation measures were promising, whereas the estimations of likelihood and severity were not valuable.

Beyond the technical results, we also acknowledge more general concerns with LLM-based applications. For the risk of AI perpetuating biases, it sounds ironic if not misplaced to assess the impact of AI through AI. Yet, particularly in preliminary stages of conception, when only product-owners and developers are involved, a tool like AFRIA may support pre-screening, counterbalancing primarily value-driven attitudes, and building upon broader knowledge. In this sense, one may even envision the development of dedicated LLMs finetuned over known problematic cases of AI application. In this perspective, besides continuing experimenting with other LLMs, applying more effective techniques like RAG and DSPy, and adding components in the pipeline (eg. for automated annotation and better estimations of likelihood and severity), further research should explore how automated impact assessment can be implemented responsibly (e.g. at which stage of the assessment process), and doctrinal research could help regulating this usage, e.g. clarifying the types of harms, the harm dimensions, and the aspects of fundamental rights to be considered for the assessment.

References

[1] Skoric V, Sileno G, Ghebreab S. Roles of Standardised Criteria in Assessing Societal Impact of AI. In: 2024 IEEE Conference on Artificial Intelligence (CAI). IEEE; 2024. p. 1240-5.

[2] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and machines. 2018;28:689-707.

[3] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. Impact Assessment Mensenrechten en Algoritmes;. Available from: `https://open.overheid.nl/documenten/ronl-c3d7fe94-9c62-493f-b858-f56b5e246a94/pdf`.

[4] Buçinca Z, Pham CM, Jakesch M, Ribeiro MT, Olteanu A, Amershi S. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms; 2023. Available from: `https://arxiv.org/abs/2306.03280`.

[5] Binns R. Data protection impact assessments: a meta-regulatory approach. International Data Privacy Law. 2017 04;7(1):22-35.

[6] Mantelero A. The Ai Act's Fundamental Rights Impact Assessment. Available at SSRN 4782126. 2024.

[7] European Center for Not-for-Profit Law, Access Now. Towards Meaningful Fundamental Rights Impact Assessments under the DSA;. Available from: `https://www.accessnow.org/wp-content/uploads/2023/09/DSA-FRIA-joint-policy-paper-September-2023.pdf`.

[8] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al.. A Survey on Evaluation of Large Language Models. New York, NY, USA: Association for Computing Machinery; 2024.

[9] Wei C, Wang YC, Wang B, Kuo CH. An Overview of Language Models: Recent Developments and Outlook. APSIPA Transactions on Signal and Information Processing. 2024;13(2).

[10] Dziri N, Milton S, Yu M, Zaiane O, Reddy S. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? In: Carpuat M, de Marneffe MC, Meza Ruiz IV, editors. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics; 2022. p. 5271-85.

[11] Kambhampati S. Can large language models reason and plan? Annals of the New York Academy of Sciences. 2024;1534(1):15-8.

[12] Ferrara E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci. 2023;6(1):3.

[13] Esiobu D, Tan X, Hosseini S, Ung M, Zhang Y, Fernandes J, et al. ROBBIE: Robust bias evaluation of large generative language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 3764-814.

[14] Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. Human decisions and machine predictions. The quarterly journal of economics. 2018;133(1):237-93.

[15] Ribeiro MT, Lundberg S. Adaptive Testing and Debugging of NLP Models. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 3253-67.

[16] Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J, et al. Red teaming language models with language models. arXiv preprint arXiv:220203286. 2022.

[17] Constantinides M, Bogucka EP, Scepanovic S, Quercia D. Good Intentions, Risky Inventions: A Method for Assessing the Risks and Benefits of AI in Mobile and Wearable Uses. Proc ACM Hum-Comput Interact. 2024 Sep;8(MHCI).

[18] Kamil MZ, Taleb-Berrouane M, Khan F, Amyotte P, Ahmed S. Textual data transformations using natural language processing for risk assessment. Risk analysis. 2023;43(10):2033-52.

[19] Costa V, Coelho P, Castelli M. Artificial Intelligence for Impact Assessment of Administrative Burdens. Emerging Science Journal. 2024;8(1):270-82.

[20] Abdi H. Holm's sequential Bonferroni procedure. Encyclopedia of research design. 2010;1(8):1-8.