Automating Fundamental Right Impact Assessment: an Open Experiment

JURIX conference, Brno, 13 December, 2024

Xinyue Zhang, Vanja Skoric, Giovanni Sileno
University of Amsterdam

Challenge

- AI applications are increasingly developed and deployed in human activities, in the public, private, and civic sector

- AI Act art 27: **FRIA** (Fundamental Right Impact Assessment)
  - requires: 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
- Additional recommendation by scholars
  - focus on impacted fundamental rights and risk and likelihood
- Resource-intensive
  - disadvantages small-medium institutions

Challenge

- AI applications are increasingly developed and deployed in human activities, in the public, private, and civic sector

- AI Act art 27: **FRIA** (Fundamental Right Impact Assessment)
    - requires: 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
- Additional recommendation by scholars
    - focus on impacted fundamental rights and risk and likelihood

expensive, and required?

urgent to facilitate FRIA, particularly at the early stages of development

Research relevance

- Few solutions proposed in support so far.
- The most relevant AHA! (2023) [Harvard, Microsoft, …], but closed-source

arXiv > cs > arXiv:2306.03280

Search...

Help | A

Computer Science > Human-Computer Interaction

[Submitted on 5 Jun 2023]

AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms

Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, Saleema Amershi

While demands for change and accountability for harmful AI consequences mount, foreseeing the downstream effects of deploying AI systems remains a challenging task. We developed AHA! (Anticipating Harms of AI), a generative framework to assist AI practitioners and decision-makers in anticipating potential harms and unintended consequences of AI systems prior to development or deployment. Given an AI deployment scenario, AHA! generates descriptions of possible harms for different stakeholders. To do so, AHA! systematically considers the interplay between common problematic AI behaviors as well as their potential impacts on different stakeholders, and narrates these conditions through vignettes. These vignettes are then filled in with descriptions of possible harms by prompting crowd workers and large language models. By examining 4113 harms surfaced by AHA! for five different AI deployment scenarios, we found that AHA! generates meaningful examples of harms, with different problematic AI behaviors resulting in different types of harms. Prompting both crowds and a large language model with the vignettes resulted in more diverse examples of harms than those generated by either the crowd or the model alone. To gauge AHA!'s potential practical utility, we also conducted semi-structured interviews with responsible AI professionals (N=9). Participants found AHA!'s systematic approach to surfacing harms important for ethical reflection and discovered meaningful stakeholders and harms they believed they would not have thought of otherwise. Participants, however, differed in their opinions about whether AHA! should be used upfront or as a secondary-check and noted that AHA! may shift harm anticipation from an ideation problem to a potentially demanding review problem. Drawing on our results, we discuss design implications of building tools to help practitioners envision possible harms.

Method for AFRIA *(Automated Fundamental Rights Impact Assessment)*

- **Intuition:** Use LLMs as information retrieval tools. If they capture common-sense knowledge, if adequately prompted they should help identifying expected outcomes of certain scenarios.

Method for AFRIA *(Automated Fundamental Rights Impact Assessment)*

- **Intuition:** Use LLMs as information retrieval tools. If they capture common-sense knowledge, if adequately prompted they should help identifying expected outcomes of certain scenarios.
- Prompting can be iterative, *to keep control of the chaining of inferences*.

Method for AFRIA *(Automated Fundamental Rights Impact Assessment)*

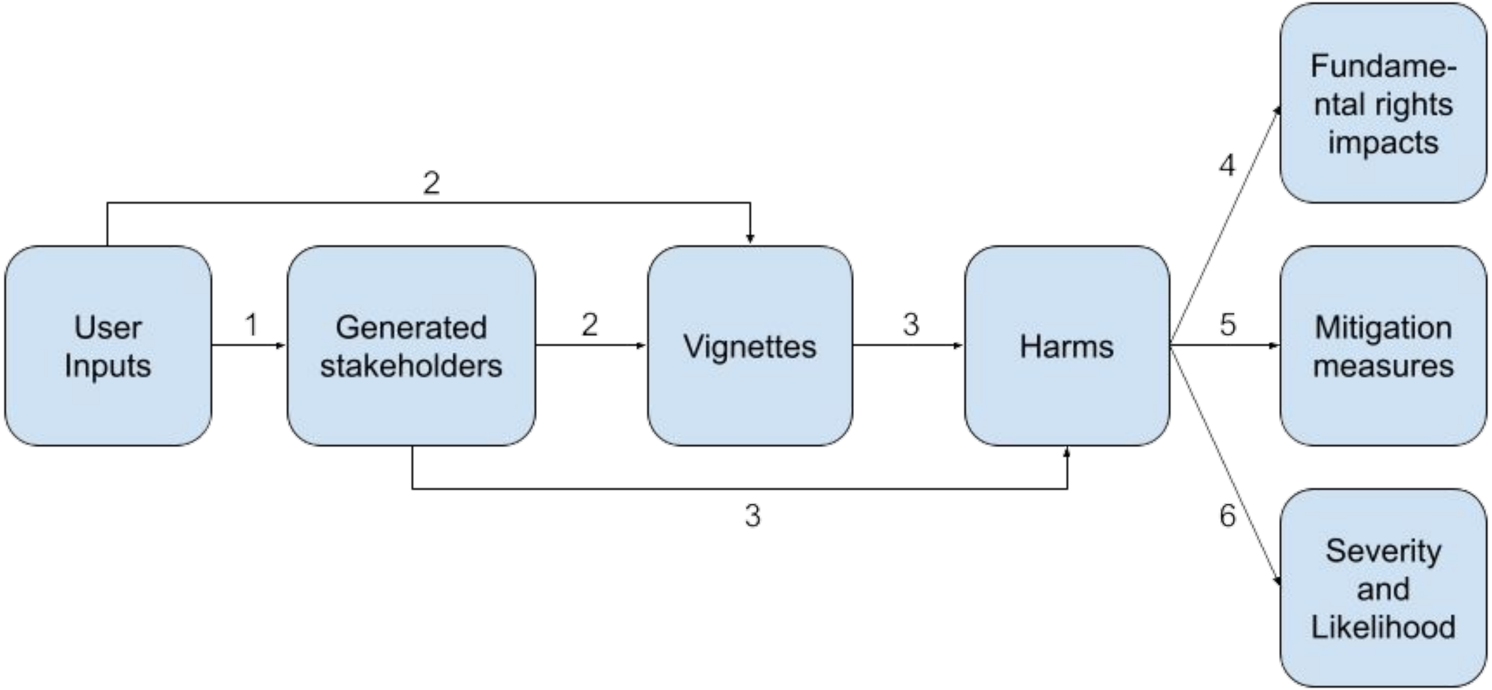
- **Intuition:** Use LLMs as information retrieval tools. If they capture common-sense knowledge, if adequately prompted they should help identifying expected outcomes of certain scenarios.
- Prompting can be iterative, *to keep control of the chaining of inferences*.
- We follow the general schema applied by AHA!, yet to be extended encompassing critics from legal scholars, like:
    - focus on the fundamental human rights impacted
    - mitigation measures
    - consider severity and likelihood

Research questions

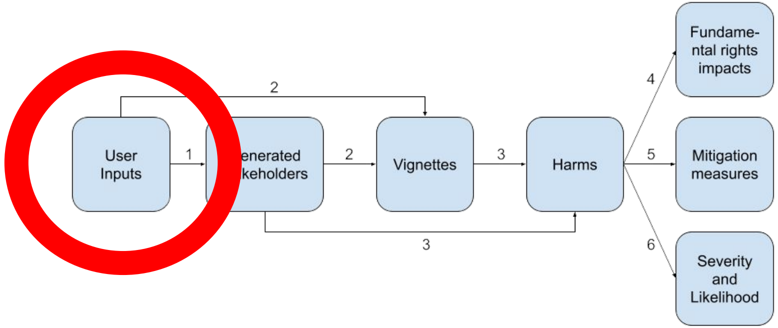
- RQ1: Can AFRIA generate meaningful examples of harms?
- RQ2: Do categories of harms differ significantly depending on the scenario?
- RQ3: Do categories of harms differ significantly depending on the dimension of problematic AI behavior?
- RQ4: Can AFRIA generate meaningful examples of fundamental rights impacts?

AFRIA Pipeline

Automated Fundamental Rights Impact Assessment

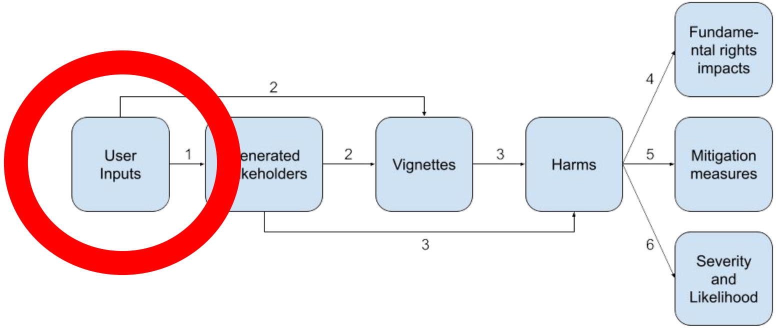


Step 0: User input



1. Scenario
2. Relevant stakeholders
3. Harm dimensions (type of harms, e.g. due to ***false positive*** errors)
    a. (Optional) specific harms (harm that results in e.g. financial strain)

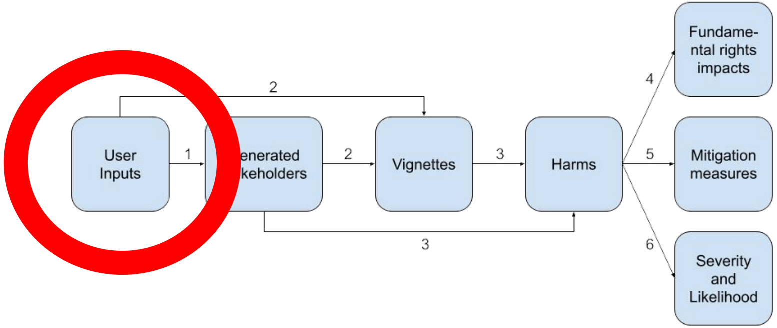
Step 0: User input



1. **Scenario**
2. Relevant stakeholders
3. Harm dimensions (type of harms, e.g. due to ***false positive*** errors)
    a. (Optional) specific harms (harm that results in e.g. financial strain)

scenario: "**A tech company wants to deploy an AI hiring system to scan the resumes of applicants and predict whether they are a good fit for a given job opening.**"

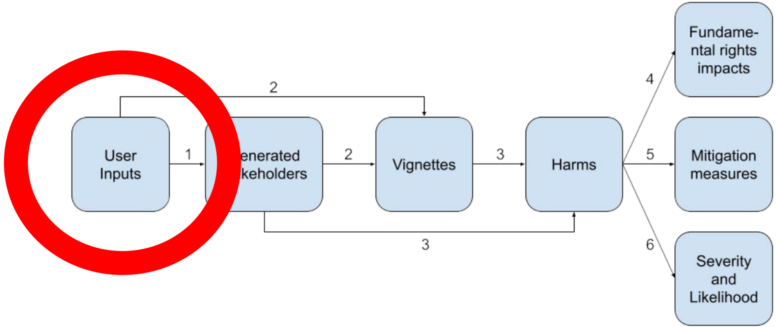
Step 0: User input



1. Scenario
2. **Relevant stakeholders**
3. Harm dimensions (type of harms, e.g. due to ***false positive*** errors)
    a. (Optional) specific harms (harm that results in e.g. financial strain)

stakeholders: "**the applicant**", "other applicants", "future applicants", "the hiring manager", "the HR team", "the company", "the AI system developers", "the family/friends of the applicant", "the applicants who identify as racial or ethnic minorities", "the applicants who identify as women", "the applicants who identify both as women and as racial or ethnic minorities”

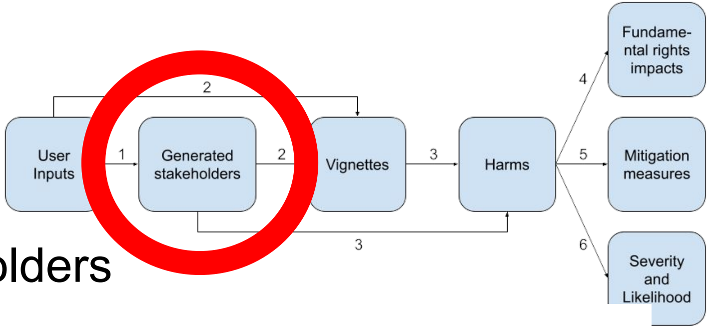
Step 0: User input



1. Scenario
2. Relevant stakeholders
3. **Harm dimensions** (type of harms, e.g. due to ***false positive*** errors)
   a. (Optional) specific harms (harm that results in e.g. financial strain)

harm dimensions: "**false positives (when the system predicts an applicant is suitable while they are not)**", "false negatives (when the system predicts an applicant is not suitable while they are)", "one-time false-positive (when the system makes a one-time mistake of predicting that an applicant is suitable while they are not)", ….

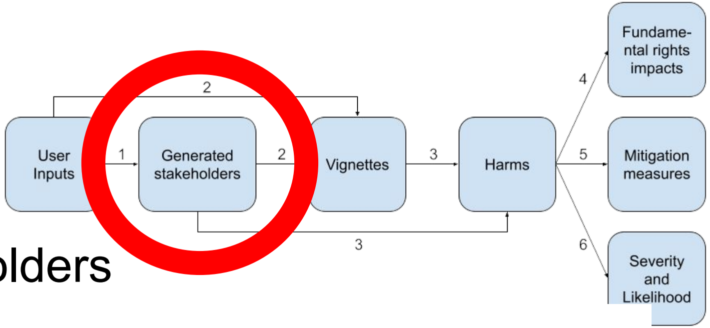
Step 1: Generate stakeholders



Using the inputs, generate another list of stakeholders

- Broader list
- Filters irrelevant stakeholders

Step 1: Generate stakeholders



Using the inputs, generate another list of stakeholders

- Broader list
- Filters irrelevant stakeholders

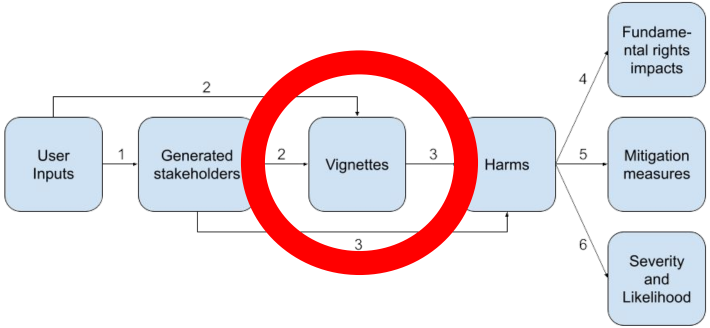
Direct Stakeholders
1. The tech company deploying the AI hiring system
2. The hiring manager using the AI system to review resumes
3. The HR team responsible for overseeing the hiring process
4. The applicants whose resumes are being scanned by the AI system
5. The AI system developers who created and maintain the technology

Indirect Stakeholders:
1. Other applicants who may be impacted by the AI system's decisions
2. Future applicants who may be subject to the same AI screening process
3. Family and friends of the applicants who may be affected by the outcome of the hiring process
4. Applicants who identify as racial or ethnic minorities, as the AI system may have biases that influence their chances of being hired
5. Applicants who identify as women, as the AI system may have biases that affect their opportunities in the hiring process
6. Applicants who identify as both women and as racial or ethnic minorities, as they may face intersectional biases from the AI system.

Step 2: Generate vignette

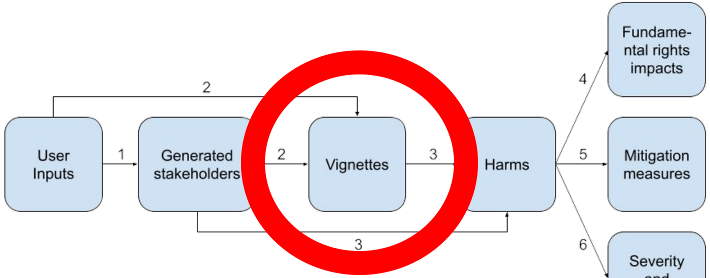
vignette = harmful scenario

For each stakeholder, generate relevant harmful scenarios for each harm dimension (FP, FN, …)



Step 2: Generate vignette

vignette = harmful scenario



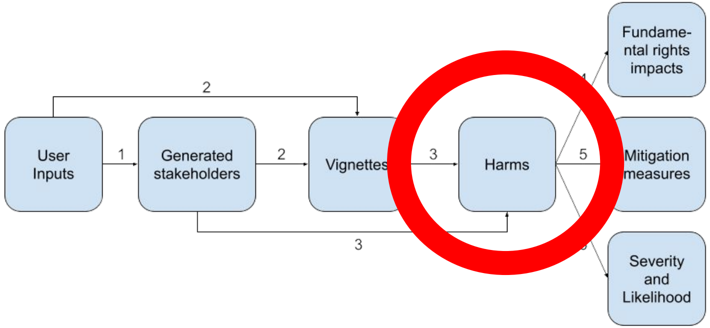
For each stakeholder, generate relevant harmful scenarios for each harm dimension (FP, FN, …)

Stakeholder: The applicant
Harm dimension: false positive (FP)

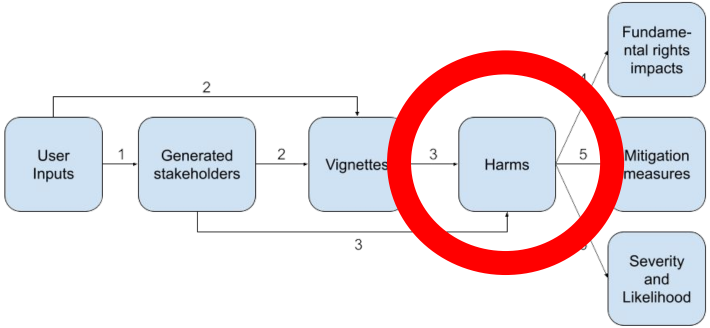
Vignette: "Imagine you are an applicant whose resume is being scanned by the AI system. You have worked hard on crafting a strong resume that highlights your relevant skills and experiences for the job you are applying to. However**, due to false positives generated by the AI system, the system mistakenly predicts that you are a good fit for the job when in reality you may not have the required qualifications or experience**. As a result, you may be selected for further rounds of interviews or assessments based on the inaccurate assessment by the AI system. This can lead to wasted time and effort on both your end and the company's end as you may not ultimately be the right candidate for the job. Additionally, false positives can also result in mismatched job placements, leading to potential dissatisfaction and underperformance in the role for which you were mistakenly selected."

Step 3: Generate harms

For each vignette, specify the harm.



Step 3: Generate harms

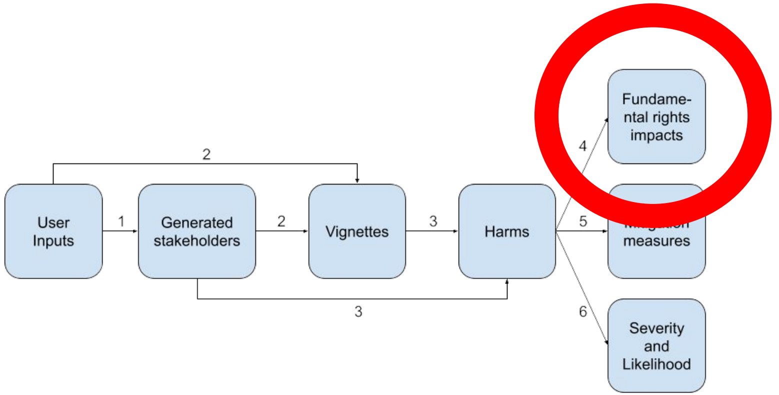


For each vignette, specify the harm.

harm matrix cell

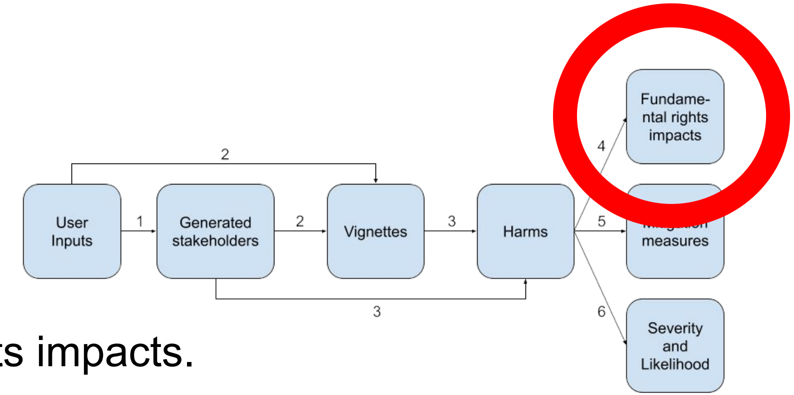
Scenario 1: Hiring ("A tech company … job opening")						
	false positive (FP)	**False Negative (FN)**	**One-time FP**	**One-time FN**	**Accumulated FP**	**…**
The applicant	"Imagine you are an applicant whose resume is being scanned by the AI system. Due to false positives generated by the system, you might be mistakenly selected for further interview rounds or assessments, even though you may not have the necessary qualifications or experience for the job. This can result in wasted time and effort for both you and the company. Additionally, being mismatched for a job can lead to dissatisfaction and underperformance in a role that you were inaccurately selected for, ultimately affecting your career prospects and job satisfaction."	"Imagine you are an applicant applying for a job at a tech company. Despite having relevant experience and skills for the position, the AI hiring system mistakenly categorizes your resume as not suitable for the job. This false negative harms you by not considering you for the position, resulting in you missing out on the opportunity to showcase your qualifications and potential to the hiring team. This unfair disadvantage could hinder your chances of advancing in the recruitment process compared to other applicants who were not overlooked by the faulty algorithm, impacting your career opportunities and advancement prospects."	…	…	…	…
…	…	…	…	…	…	…

Step 4: Generate Fundamental Rights Impacted



For each harm, specify the fundamental rights impacts.

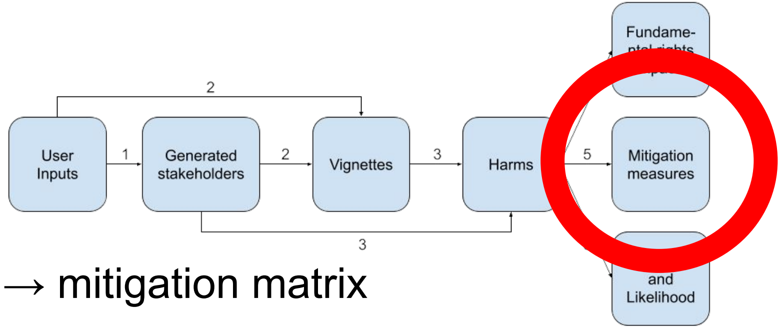
Step 4: Generate Fundamental Rights Impacted



For each vignette, harm, specify the fundamental rights impacts.

Scenario 1: Hiring ("A tech company … job opening")			
	false positive (FP)	**False Negative (FN)**	**…**
The applicant	"The universal Declaration of Human Rights enshrines several relevant rights for the applicants being evaluated by the AI system in the context described: 1. **Right to Work:** The right to work is protected under Article 23, stating that everyone has the right to work, to free choice of employment, to just and favorable conditions of work, and to protection against unemployment. In the case of AI systems evaluating job applicants, a false positive could impact this right by misleading the applicant about their actual qualifications for the job. 2**. Right to Fair Treatment in Employment:** Article 23 of the Declaration also includes the right to equal pay for equal work and the right to just and favorable conditions of work. If an AI system mistakenly categorizes an applicant as a strong fit for a job when they are not, it could lead to unfair treatment in employment by setting false expectations for the applicant. **3. Right to Education**: The right to education is recognized in Article 26, stating that everyone has the right to education."	The human rights of the applicants being evaluated by the AI system that are affected by this harm include: **1. Right to Work:** The right to work is a fundamental human right that emphasizes the opportunity for everyone to freely choose their work and employment. When AI systems introduce bias or errors in the evaluation process, it can unfairly hinder an applicant's chances of securing employment that they are qualified for, thus infringing on their right to work. **2. Right to Non-Discrimination:** The right to non-discrimination ensures that individuals are treated fairly and equally in all aspects, including employment opportunities. If AI systems exhibit bias that leads to false negative evaluations based on irrelevant factors such as race, gender, or socio-economic background, it violates the applicants' right to be free from discrimination. **3. Right to Education and Training:** The right to education and training is essential for individuals to acquire the skills and knowledge necessary to pursue their chosen careers. When AI systems inaccurately assess applicants, qualified candidates may be overlooked,	…
…	…	…	…

Step 5: Generate Mitigation Measures [exploratory]

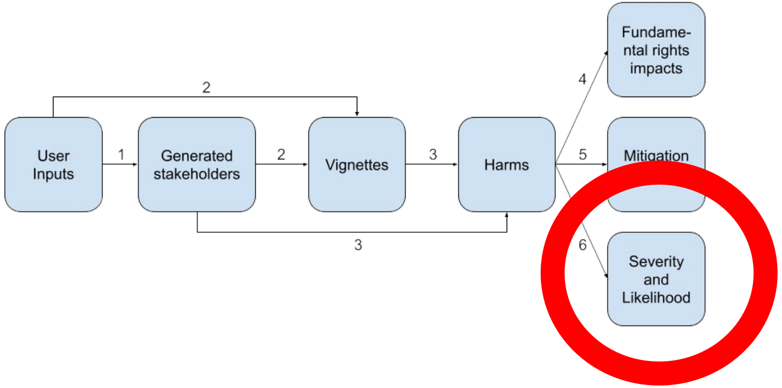


For each potential harm, how harm may be mitigated? → mitigation matrix

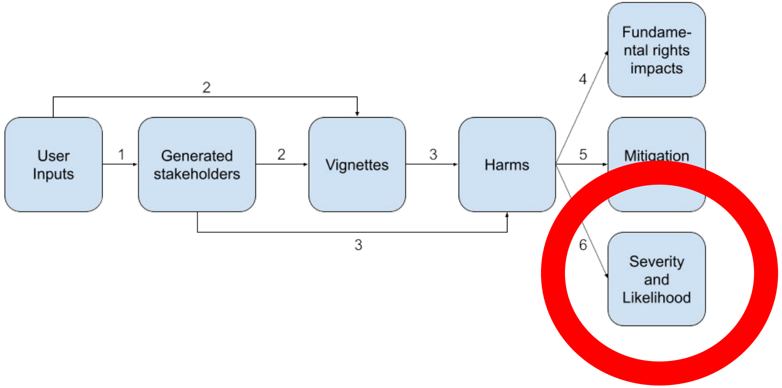
Step 6: Generate Severity and Likelihood [exploratory]

For each harm,

- what is its severity? → severity matrix
- what is its likelihood? → likelihood matrix



Step 6: Generate Severity and Likelihood [exploratory]



For each harm,

- what is its severity?
- what is its likelihood?
- how much are you confident in this estimation?

meta-cognition task

Evaluation: baseline

We consider the same application context presented in Buçinca et al. (2023).
***AHA**!: Facilitating ai impact assessment by generating examples of harms.*

- GPT 3.5-turbo
- 5 scenarios (communication compliance, hiring, loan application, etc.)
- List of inputs (scenarios, stakeholders, harm dimensions)
- Taxonomy of harms

Baseline Model (AHA!)	**AFRIA Model (AHA!)**
4113 harms	4580 (after “splitting” harm matrix cells) harms
No FR	402 (after “splitting” FR matrix cells) FR

Meaningful vs Sensical vs Nonsensical

We apply the same terminology used in AHA!, but clarify it further:

- **Meaningful**:
  - (1) the connection between action and consequences is plausible or logical;
  - (2) the consequences have actual harmful effects;
  - (3) the harm is affecting the target stakeholder.
- Non-meaningful consists of
  - **Nonsensical**: absence of criteria (1)
  - **Sensical**: presence of criteria (1), but absence of (2) and/or (3)

Evaluation and statistical analysis (Harm)

- RQ1: percentage meaningful harms 93.7%
- RQ2:
  - manual categorisation based on AHA! harm taxonomy
  - chi-square analysis with Holm-Bonferroni correction to analyse the significance of the distribution

		scenario					
		Communication compliance	Content moderation	Disease diagnosis	Hiring	Loan application	Total
What is the harm category?	Quality of service harms	63	74	89	94	48	368
	Representional harms	31	81	4	48	31	195
	Well-being harms	239	272	212	171	96	990
	Legal and Reputational harms	270	264	226	159	155	1074
	Social and societal harms	244	100	25	53	46	468
	Loss of rights or agency	35	81	2	0	1	119
	Allocational harms	175	115	174	307	259	1030
	Other harms	3	4	2	4	2	15
	Not meaningful	22	12	10	22	54	120
Total		1082	1003	744	858	692	4379

Evaluation and statistical analysis (Harm)

- RQ1: percentage meaningful harms 93.7%
- RQ2:
  - manual categorisation based on AHA! harm taxonomies
  - chi-square analysis with Holm-Bonferroni correction to analyse the significance of the distribution
- RQ 3:
  - chi-square analysis to analyse the significant of the category distribution per harm dimensions pair

	False positive / negative				Accumulated / one-time				Egregious / not specified			
Scenario	**N**	**df**	**chi-square**	**p-value**	**N**	**df**	**chi-square**	**p-value**	**N**	**df**	**chi-square**	**p-value**
Communication compliance	173	7	2,554	0,923 (n.s.)	349	8	8,312	0,404 (n.s.)	389	8	4,67	0,792 (n.s.)
Content moderation	167	7	14,331	**0,046**	348	8	5,873	0,661 (n.s.)	342	8	2,329	0,969 (n.s.)
Disease diagnosis	127	8	10,989	0,202 (n.s.)	253	7	3,715	0,812 (n.s.)	252	7	6,356	0,510 (n.s.)
Hiring	144	6	3,58	0,733 (n.s.)	306	6	2,257	0,895 (n.s.)	291	7	6,447	0,489 (n.s.)
Loan application	116	6	1,47	0,961 (n.s.)	234	7	6,804	0,450 (n.s.)	246	7	5,799	0,563 (n.s.)

Evaluation and statistical analysis (Harm)

- RQ1: percentage meaningful harms 93.7%
- RQ2:
  - manual categorisation based on AHA! harm taxonomies
  - chi-square analysis with Holm-Bonferroni correction to analyse the significance of the distribution
- RQ 3:
  - chi-square analysis to ana
  - heat map and chi-square for specified harms

Count

		scenario					
		Communication compliance	Content moderation	Disease diagnosis	Hiring	Loan application	Total
What is the harm category?	Quality of service harms	11	6	17	6	4	44
	Representional harms	4	13	0	2	2	21
	Well-being harms	64	62	38	16	11	191
	Legal and Reputational harms	34	34	29	16	14	127
	Social and societal harms	32	8	0	4	4	48
	Loss of rights or agency	2	11	0	0	0	13
	Allocational harms	23	10	24	67	56	180
	Other harms	0	0	0	0	1	1
	Not meaningful	1	2	3	6	4	16
Total		171	146	111	117	96	641

Evaluation and statistical analysis (FR)

- RQ4: percentage meaningful FR 58.2%
  - manually label the FR impacts
  - manual categorisation into the original FR impacts taxonomy

FR_categories * scenario Crosstabulation

Count

		scenario					
		Communication compliance	Content moderation	Disease diagnosis	Hiring	Loan application	Total
FR_categories	Work-related FR	17	0	4	22	28	71
	Finance-related FR	0	0	0	0	13	13
	Non-discrimination FR	20	7	9	25	11	72
	Remedy-related FR	2	2	1	11	10	26
	Expression-related FR	17	17	10	1	6	51
	Well-being-related FR	1	3	18	3	9	34
	Privacy-related FR	21	14	9	28	3	75
Total		78	43	51	90	80	342

Results on Harms

Baseline (AHA!)	**AFRIA**
93% meaningful harms	93.7% meaningful harms
7% non-meaningful harms, of which 86.4% nonsensical	6.3% non-meaningful harms, of which 95.1% nonsensical
Meaningful distribution across harm categories	Meaningful distribution of harm across harm categories
Largely unmeaningful distribution across harm dimensions	Worse

Results on Fundamental Rights (FR)

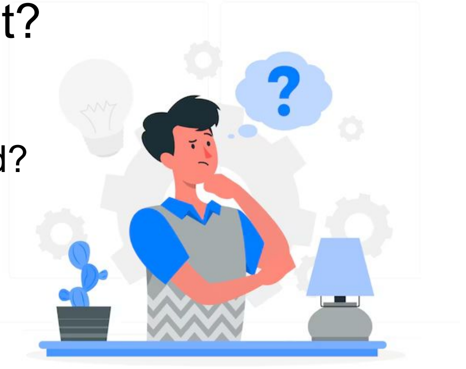
Baseline (AHA!)	AFRIA
N.A.	58.2% meaningful FR
N.A.	41.8% non-meaningful harms, of which 87.8% nonsensical
N.A.	Meaningful distribution across FR categories

Preliminary results on exploratory dimensions

- mitigation measures: promising
- severity and likelihood: always high/very high
- confidence: always very confident! (the LLM visibly fails the meta-cognition task)

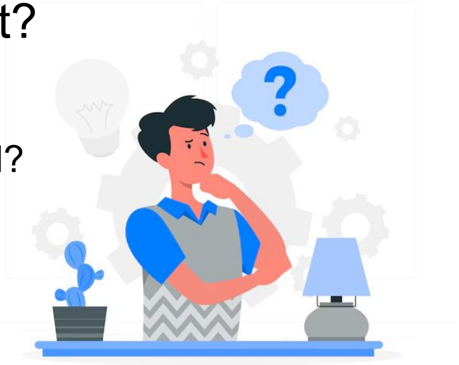
Discussion

- Are LLM the “proper” mean for AI impact assessment?
    - ethical, environmental, privacy concerns
    - risk to remove the human component where it is more needed?



Discussion

- Are LLM the “proper” mean for AI impact assessment?
  - ethical, environmental, privacy concerns
  - risk to remove the human component where it is more needed?



Limitations of the study

- Evaluation based on error-prone human annotation
- Simple prompting techniques
- LLM used not state of the art

Future work

- Try different LLMs (open-source) and other methods (RAG, DSPy, etc.)
- Consider automated classification instead of manual annotation
- Improve harm distribution across harm dimensions (e.g. FP/FN)
- Systematize the analysis of likelihood, severity, …

Automating Fundamental Right Impact Assessment: an Open Experiment

JURIX conference, Brno, 13 December, 2024

Xinyue Zhang, Vanja Skoric, Giovanni Sileno
University of Amsterdam

Example

1. Input
    a. scenario: "A tech company wants to deploy an AI hiring system to scan the resumes of applicants and predict whether they are a good fit for a given job opening."
    b. stakeholders: "the applicant", "other applicants", "future applicants", "the hiring manager", "the HR team", "the company", "the AI system developers", "the family/friends of the applicant", "the applicants who identify as racial or ethnic minorities", "the applicants who identify as women", "the applicants who identify both as women and as racial or ethnic minorities"
    c. harm dimensions: "False positives (when the system predicts an applicant is suitable while they are not)", "false negatives (when the system predicts an applicant is not suitable while they are)", "one-time false-positive (when the system makes a one-time mistake of predicting that an applicant is suitable while they are not)", "accumulated false-positive (when the system repeatedly or systematically over time during its deployment predicts an applicant is suitable while they are not)", "one-time false-negative (when the system makes a one-time mistake and predicts an applicant is not suitable while they are)", "accumulated false-negative (when the system repeatedly or systematically over time during its deployment predicts an applicant is not suitable while they are)", "egregious false positives (when the system makes a severe error and predicts an applicant is suitable while they are not)", "unspecified false positives (when the system makes an error of unspecified severity and predicts an applicant is suitable while they are not)", "egregious false negatives (when the system makes a severe error and predicts an applicant is not suitable while they are)", "unspecified false negatives (when the system makes an error of unspecified severity and predicts a predicts an applicant is not suitable while they are)", "financial strain because the system predicts an applicant is suitable while they are not", "financial strain because the system predicts an applicant is not suitable while they are"

Example results

2. Stakeholders:

```
"Direct Stakeholders:",

   "1. The tech company deploying the AI hiring system",
   "2. The hiring manager using the AI system to review resumes",
   "3. The HR team responsible for overseeing the hiring process",
   "4. The applicants whose resumes are being scanned by the AI system",
   "5. The AI system developers who created and maintain the technology",
   "",
   "Indirect Stakeholders:",
   "1. Other applicants who may be impacted by the AI system's decisions",
   "2. Future applicants who may be subject to the same AI screening process",
   "3. Family and friends of the applicants who may be affected by the outcome of the hiring process",
   "4. Applicants who identify as racial or ethnic minorities, as the AI system may have biases that influence their chances of being hired",
   "5. Applicants who identify as women, as the AI system may have biases that affect their opportunities in the hiring process",
   "6. Applicants who identify as both women and as racial or ethnic minorities, as they may face intersectional biases from the AI system."
```

Example results

3. Vignette

```
The applicants whose resumes are being scanned by the AI system": {
       "False positives (when the system predicts an applicant is suitable while they are not)":
"Imagine you are an applicant whose resume is being scanned by the AI system. You have worked hard on
crafting a strong resume that highlights your relevant skills and experiences for the job you are
applying to. However, due to false positives generated by the AI system, the system mistakenly predicts
that you are a good fit for the job when in reality you may not have the required qualifications or
experience.\n\nAs a result, you may be selected for further rounds of interviews or assessments based on
the inaccurate assessment by the AI system. This can lead to wasted time and effort on both your end and
the company's end as you may not ultimately be the right candidate for the job. Additionally, false
positives can also result in mismatched job placements, leading to potential dissatisfaction and
underperformance in the role for which you were mistakenly selected.",
       "false negatives (when the system predicts an applicant is not suitable while they are)":
"Imagine
```

Example results

3. Harm:

```
The applicants whose resumes are being scanned by the AI system": {

       "False positives (when the system predicts an applicant is suitable while they are not)": "Imagine you are an applicant whose resume is being scanned by the AI system. Due to false positives generated by the system, you might be mistakenly selected for further interview rounds or assessments, even though you may not have the necessary qualifications or experience for the job. This can result in wasted time and effort for both you and the company. Additionally, being mismatched for a job can lead to dissatisfaction and underperformance in a role that you were inaccurately selected for, ultimately affecting your career prospects and job satisfaction.",
       "false negatives (when the system predicts an applicant is not suitable while they are)": "Imagine …
```

Example results

4. FR Impact:

The applicants being evaluated by the AI system": {

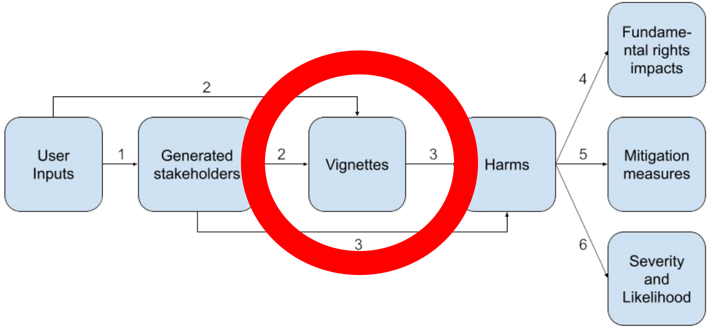
"False positives (when the system predicts an applicant is suitable while they are not)": "The universal Declaration of Human Rights enshrines several relevant rights for the applicants being evaluated by the AI system in the context described:\n\n1. Right to Work: The right to work is protected under Article 23, stating that everyone has the right to work, to free choice of employment, to just and favorable conditions of work, and to protection against unemployment. In the case of AI systems evaluating job applicants, a false positive could impact this right by misleading the applicant about their actual qualifications for the job.\n\n2. Right to Fair Treatment in Employment: Article 23 of the Declaration also includes the right to equal pay for equal work and the right to just and favorable conditions of work. If an AI system mistakenly categorizes an applicant as a strong fit for a job when they are not, it could lead to unfair treatment in employment by setting false expectations for the applicant.\n\n3. Right to Education: The right to education is recognized in Article 26, stating that everyone has the right to education.",
"false negatives (when the system predicts an applicant is not suitable while they are)": "The human rights …

Example results

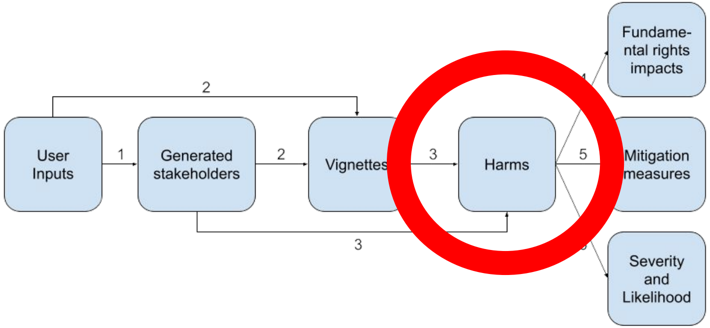
1. Input
   a. scenario:
   b. stakeholders:
   c. harm dimensions
2. Vignette:
3. Harm:
   a. meaningful:
   b. non-meaningful
4. FR impacts
   a. meaningful:
   b. non-meaningful:

Step 2: Generate vignette

Example:



Step 3: Generate harms



Evaluation #3

- [here talk of the various taxonomies]
- [and of the manual annotation]

Harm dimensions: specified harms distribution

What is the harm category? * scenario Crosstabulation

Count

		scenario					
		Communication compliance	Content moderation	Disease diagnosis	Hiring	Loan application	Total
What is the harm category?	Quality of service harms	11	6	17	6	4	44
	Representional harms	4	13	0	2	2	21
	Well-being harms	64	62	38	16	11	191
	Legal and Reputational harms	34	34	29	16	14	127
	Social and societal harms	32	8	0	4	4	48
	Loss of rights or agency	2	11	0	0	0	13
	Allocational harms	23	10	24	67	56	180
	Other harms	0	0	0	0	1	1
	Not meaningful	1	2	3	6	4	16
Total		171	146	111	117	96	641