ETHICAL IMPLICATIONS OF AI: ASSESSING TRUSTWORTHY AI IN PRACTICE

2022-02 GSDS HEEJIN KIM

HUMAN RIGHTS IMPACT ASSESSMENT (HRIA)

(as mainly derived from the *UN Guiding Principles on Business and Human Rights* and other legal documents – from the last class)

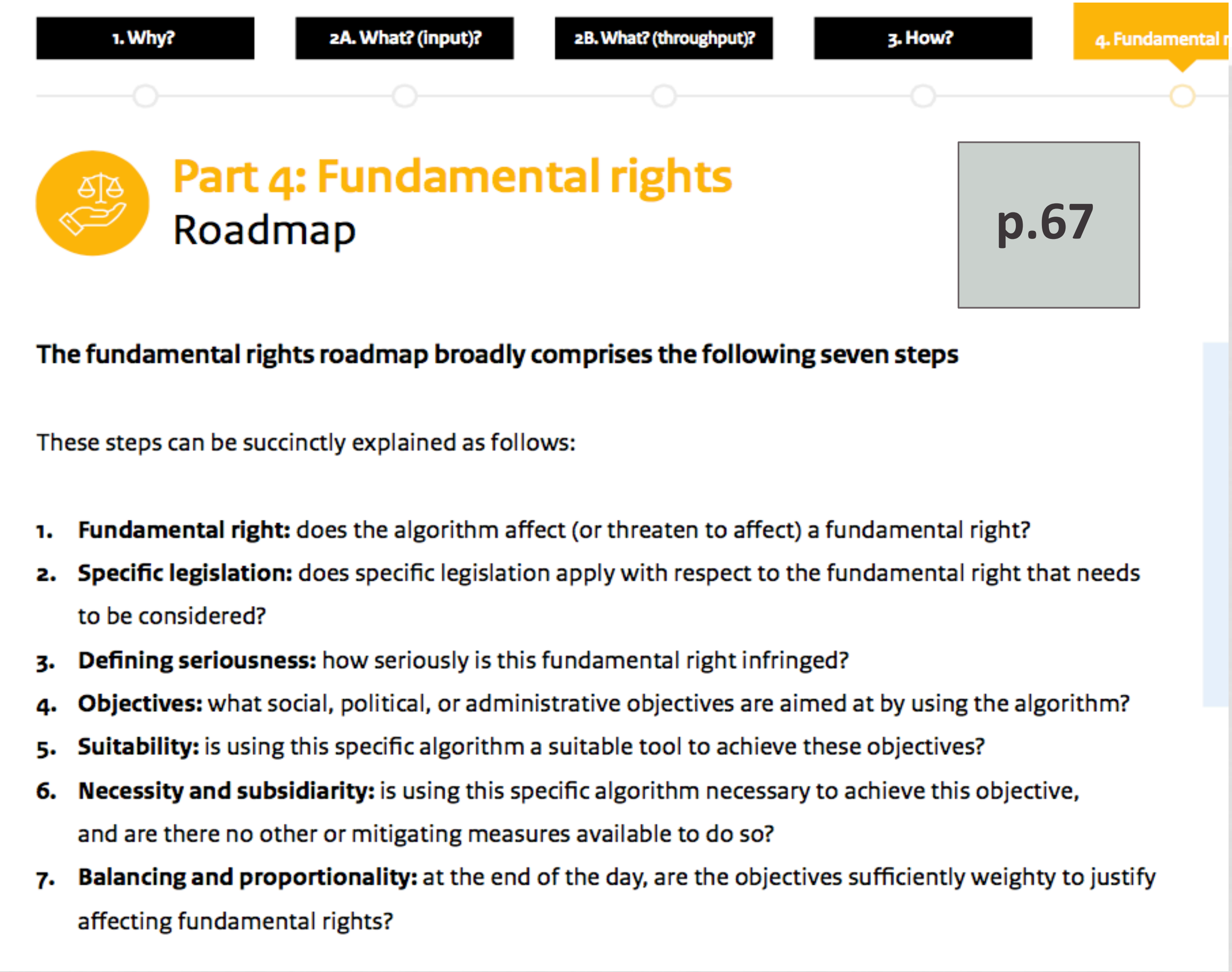
- [e.g.] **A process** for identifying, understanding, assessing & addressing impact of business project or activities on **human rights of the impacted right-holders**.
  - Impact assessment comes with different phases or steps, all of which to be used for assessment
- UN agencies, governments, companies have conducted human rights impact assessment in the tech sector – recently including AI
- [for instance, as uploaded @ course webpage)

Ministry of the Interior and
Kingdom Relations

March 2022

Impact Assessment
Fundamental rights and algorithms

HUMAN RIGHTS IMPACT ASSESSMENT



Part 4: Fundamental rights

Roadmap

p.67

The fundamental rights roadmap broadly comprises the following seven steps

These steps can be succinctly explained as follows:

1. **Fundamental right:** does the algorithm affect (or threaten to affect) a fundamental right?
2. **Specific legislation:** does specific legislation apply with respect to the fundamental right that needs to be considered?
3. **Defining seriousness:** how seriously is this fundamental right infringed?
4. **Objectives:** what social, political, or administrative objectives are aimed at by using the algorithm?
5. **Suitability:** is using this specific algorithm a suitable tool to achieve these objectives?
6. **Necessity and subsidiarity:** is using this specific algorithm necessary to achieve this objective, and are there no other or mitigating measures available to do so?
7. **Balancing and proportionality:** at the end of the day, are the objectives sufficiently weighty to justify affecting fundamental rights?

HUMAN RIGHTS IMPACT ASSESSMENT (HRIA)

Possible Roadmap for Human Rights Impact Assessment concerning your project

- [1] **Preliminary stage –** Objectives and Features
- [2] **Planning & Scoping I –** Initial Assessment
- [3] **Planning & Scoping II –** Whose rights & Whose obligations?
- [4] **Looking into the Development Process with Questions :** Data & Algorithmic Development
- [5] **Measuring –** The Impact of AI System of Your Choice & Its Seriousness
- [6] **Evaluating with Some Parameters** (necessity, subsidiarity, balancing & proportionality etc.)

HRIA FOR THE PROJECT

[1] **Preliminary stage –** Objectives and Features

- **Objectives** – [e.g.] social, political, economic, cultural, and/or administrative objectives to be achieved by using and developing the AI system of your choice
- **Features** – main features of the product, service and/or system and the context in which it will be used

Hello Barbie : (Produced by Mattel)

[1] **Preliminary stage –** Objectives and Features

- An interactive doll produced by Mattel as IoT device – equipped with speech recognition systems & AI-based learning features

[The design goal was to provide the doll with the following]

- (i) programmed with more than 8000 lines of **dialogue hosted in the cloud**, enabling the doll to talk with the user about "friends, school, dreams and fashion"
- (ii) **speech recognition technology** activated by a button on the **doll's belt buckle**
- (iii) equipped with **a microphone, speaker** and LEOs **embedded in the doll's necklace**, which light up when it is active
- (iv) a **Wi-Fi** connection to provide for **two-way conversation**

HRIA FOR THE PROJECT

[2] Planning and Scoping I – Initial Assessment

- **[A] Listing out the rights and freedoms** – potentially impacted by the AI system of your choice
- **[B] (Just roughly)** thinking about what kind of domestic, regional, international legal instruments relevant to evaluating the system (legislations and guidelines related to those rights you've listed out)

→→ [Setting the baseline where you can start]

You need to do some desk research

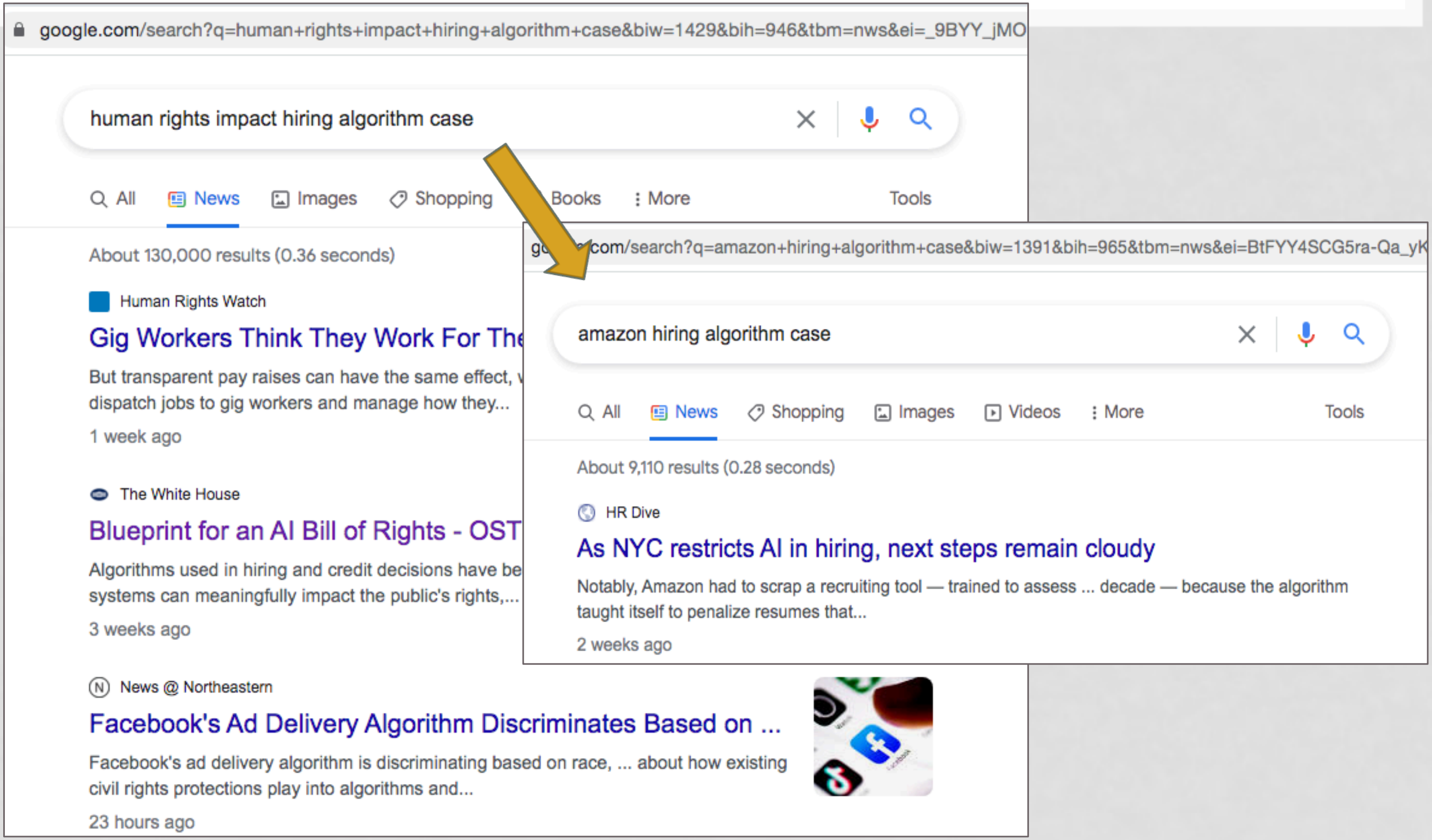
HRIA FOR THE PROJECT

[2] **Planning and Scoping I –** Initial Assessment

Desk research [e.g.]

- **Please refer to the announcement @ eTL**
  - **[e.g]** A catalogue of rights (esp p.85-95), UN Websites, the list of ICCPR; ICESCR rights from the PPT last week (substantially overlapping)
- **Type in keywords at search engines**
  - News articles, blogs, ~~academic journals~~ etc.
  - **[e.g]** just take a quick look at a couple of news that can tell you what was already said about the use of similar AI system in the same sector.

- [e.g.] Cases concerning hiring and recruitment AI system



google.com/search?q=human+rights+impact+hiring+algorithm+case&biw=1429&bih=946&tbm=nws&ei=_9BYY_jMO

human rights impact hiring algorithm case

About 130,000 results (0.36 seconds)

Human Rights Watch

Gig Workers Think They Work For The

But transparent pay raises can have the same effect, v
dispatch jobs to gig workers and manage how they...

1 week ago

The White House

Blueprint for an AI Bill of Rights - OST

Algorithms used in hiring and credit decisions have be
systems can meaningfully impact the public's rights,...

3 weeks ago

News @ Northeastern

Facebook's Ad Delivery Algorithm Discriminates Based on ...

Facebook's ad delivery algorithm is discriminating based on race, ... about how existing civil rights protections play into algorithms and...

23 hours ago

go com/search?q=amazon+hiring+algorithm+case&biw=1391&bih=965&tbm=nws&ei=BtFYY4SCG5ra-Qa_yK

amazon hiring algorithm case

About 9,110 results (0.28 seconds)

HR Dive

As NYC restricts AI in hiring, next steps remain cloudy

Notably, Amazon had to scrap a recruiting tool — trained to assess ... decade — because the algorithm taught itself to penalize resumes that...

2 weeks ago

[2] **Planning and Scoping I –** Initial Assessment

[For instance] Based on my basic understanding of the objectives & features of the system of my choice

[A] listing out the rights & freedoms potentially impacted

- IoT device heavily relying on speech recognition and data collection/use via cloud; two-way conversation
  - → ***Rights to privacy*** *and* ***personal data protection*** *may be impacted*
- Programmed with more than 8000 dialogues hosted in the cloud; talk about "friends, school, dreams and fashion" with the user → behavioral, cultural, and educational influence?
  - ***Freedom of expression, freedom of thought and diversity*** *etc.*
- Young children as main customers
  - → ***some specific rights of children?***

[2] **Planning and Scoping I –** Initial Assessment

[B] Thinking about domestic, regional, international legal instruments relevant to evaluating the system

(legislations & guidelines related to those rights you've listed out)

[e.g.] Hello Barbie case: right to privacy, personal data protection, freedom of expression, of thought, diversity, any specific right of child?

- UDHR
- ICCPR, ICESCR (core treaties)
- UN Convention on the Rights of Child (core treaties)
- UN Guiding Principle on Business and Human Rights
- Applicable domestic law

- **If uncertain, please ask me! Happy to discuss ****

HRIA FOR THE PROJECT

[3] **Planning and Scoping II –** Setting the Scene

Stakeholder issues, more precisely Q: **Whose rights and whose obligations?**

- **[A] Whose rights? – identifying the right-holders:** potential groups of people whose rights may be impacted
- Context-dependent, but here are some relevant actors
  - Target-users of the AI system of your choice?
  - Registered customers?
  - Groups of people whose data is collected and used? (general or specific)

HRIA FOR THE PROJECT

[3] Planning and Scoping II – Setting the Scene

Stakeholder issues, more precisely **Q:** Whose rights and whose obligations?

- **[B] Whose obligations? – identifying the duty-holders:**
  - Who has obligations to respect, protect and fulfill those specific rights (impacted by AI system of your choice)?
- Context-dependent, but here are some relevant actors
  - Service developers and their business partners?
  - Who own and make decisions about whether to employ a system?
  - Who are involved in the design and development? Why manage and control the system during and after deployment?

- [A] Whose rights? – **identifying the right-holders:** potential groups of people whose rights may be impacted
    - Target/direct users – most likely English-speaking child?
    - Any other groups of people?
- [B] Whose obligations? – **identifying the duty-holders:** Who has obligations to respect, protect and fulfill those specific rights (impacted by your system)?
    - Manufacturer – Mattel
    - Business partner?
    - Regulator?

- In addition: You may also want to check upon **territorial target area** [e.g.] the scope of distribution

HRIA FOR THE PROJECT

Possible Roadmap for Human Rights Impact Assessment concerning your project

- [1] **Preliminary stage –** Objectives and Features
- [2] **Planning and Scoping I –** Initial Assessment
- [3] **Planning and Scoping II –** Whose rights & Whose obligations?
- [4] **Looking into the Development Process with Questions :** Data & Algorithmic Development
- [5] **Measuring –** The Impact of AI System of Your Choice & Its Seriousness
- [6] **Evaluating with Some Parameters** (necessity, subsidiarity, balancing etc.)

HRIA FOR THE PROJECT

[4] Looking into the Development Process with Questions : Data and Algorithmic Development

Development process of the product, service, and/or system?

– Looking into the life of AI system

- Context-dependent, but here are some relevant actors
  - How is it created? How is it deployed and implemented?
  - Is it properly monitored and controlled?
- Human rights impact **stem from many sources throughout the development process** of AI system (Don’t limit yourself! Please be open-minded & creative)

[4] Looking into the Development Process with Questions : Data and Algorithmic Development

Development process of the product, service, and/or system?

- **[A] Data collection & analysis process** – “Garbage in, garbage out” problem
  - Quality of data used to train the system
  - The ways in which data are collected and processed
    - ➢ The most common approach is to look at these two and start from here. **Why?**
- **[B] algorithmic development & system design process** (context-dependent)
  - [e.g.] Certain variable you wanted AI system to optimize? What were the variables you wanted AI system to take into consideration as it operates?

[4] Looking into the Development Process with Questions : Data and Algorithmic Development

[A] Data collection and analysis process

[e.g. 1] Facial recognition technology using AI-based system

- [feature of the process: data scientists and AI researchers collect, curate and label face images] --- you may ask, for instance

  - Where & how do they get those images from?
  - From their own source of user data? Or through a commercial contract with other firms that have such data etc.
  - Subjects' awareness and/or consent? (depending on law);
  - What happens after data collection?

[4] Looking into the Development Process with Questions : Data and Algorithmic Development

[A] Data collection and analysis process

[e.g.2] Hello Barbie

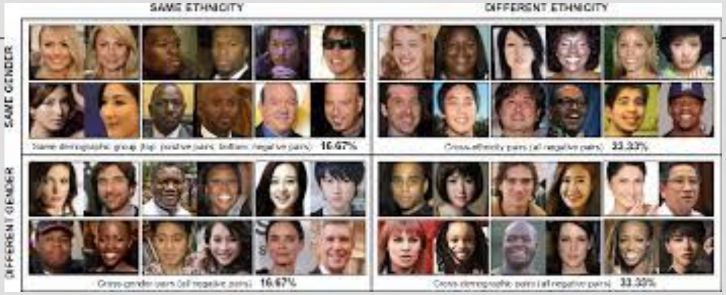
- **[feature of the process:** data to be collected via speech recognition and conversation recording] --- you may ask, for instance
  - Recording of dialogues between the doll and the users – awareness/ consent/notification?
  - Any third parties are involved in data processing?
  - **What else?**

[4] Looking into the Development Process with Questions : Data and Algorithmic Development

[B] algorithmic development & system design process

[e.g. 1] Facial recognition technology using AI-based system

- [feature of the process: from training by using the collected face data to developing general facial recognition model] --- you may ask, for instance
  - Different accuracy for different members of demographics?
  - If yes, then you connect the dot.


SAME ETHNICITY
DIFFERENT ETHNICITY
SAME GENDER
DIFFERENT GENDER
Same demographic group (top: positive pairs, bottom: negative pairs) 16.67%
Cross-ethnicity pairs (all negative pairs) 33.33%
Cross-gender pairs (all negative pairs) 16.67%
Cross-demographic pairs (all negative pairs) 33.33%

HRIA FOR THE PROJECT

[5] **Measuring –** The Impact of AI System of Your Choice & Its Seriousness

- [A] Please **try to consider** both **positive and negative** impact on human rights
- [B] Please note that the impact may not be evenly distributed across the society as well as for different groups of people
- [C] If negative impact, please consider the seriousness of the impact

[5] **Measuring –** The Impact of AI System of Your Choice & Its Seriousness

[B] Impact may not be evenly distributed across the society & not be the same for different groups of people

- [e.g.] Using **AI-based system in the context of humanitarian aid and missions** (refugees fleeing war and oppression from their home states)
  - Identifying information (e.g. names, hometowns & address)

- [e.g.] Using **facial recognition system in the public place** like airport
  - Persons with disabilities (e.g. Down’s syndrome, achondroplasia, cleft lip or palate, or other conditions that result in facial differences

[5] **Measuring –** The Impact of AI System of Your Choice & Its Seriousness

- [A] both positive and negative impact on human rights [B] the impact may not be evenly distributed across the society as well as for different groups

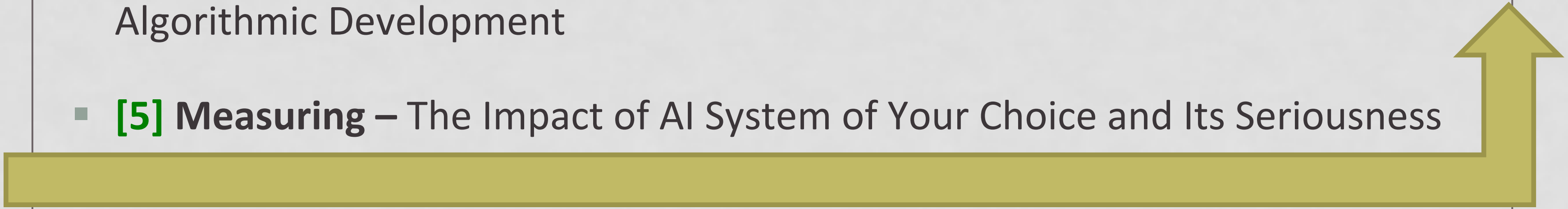
- **[C] When considering negative impact, please try to measure the seriousness of the impact**
  - ✓ [Likelihood] probability – the expected consequences on HR might occur?
  - ✓ [Actual seriousness] gravity of the consequences; and possibility of overcoming those impact?

- **→ If highly likely to happen & amount to serious violation, greater reasons to provide stronger justifications**

HRIA FOR THE PROJECT

Possible Roadmap for Human Rights Impact Assessment concerning your project

- **[1] Preliminary stage –** Objectives and Features
- **[2] Planning and Scoping I –** Initial Assessment
- **[3] Planning and Scoping II –** Whose rights & Whose obligations?
- **[4] Looking into the Development Process with Questions :** Data and Algorithmic Development
- **[5] Measuring –** The Impact of AI System of Your Choice and Its Seriousness



- **[6] Evaluating with Some Parameters** (necessity, subsidiarity, balancing etc.)

[6] Evaluating with Some Parameters

- **Compile** what you have found from phase 1 to 5 & **make a comprehensive assessment** AND **Evaluate the impact with some parameters** (+ any recommendations to respond to negative impact?)

[e.g.] balancing & proportionality – more relevant to examining government agency's obligation to protect the rights (impacted by AI system of your choice)

- **[e.g.1] Necessity**: whether the deploying the algorithm is the best option, given its impact on human rights
- **[e.g.2] Subsidiarity**: when choosing between alternatives (i.e. btw/ several types of algorithms or algorithm providers), is this the best available techniques?
- **[e.g.3] Balancing** human rights at stake with the public interests pursued