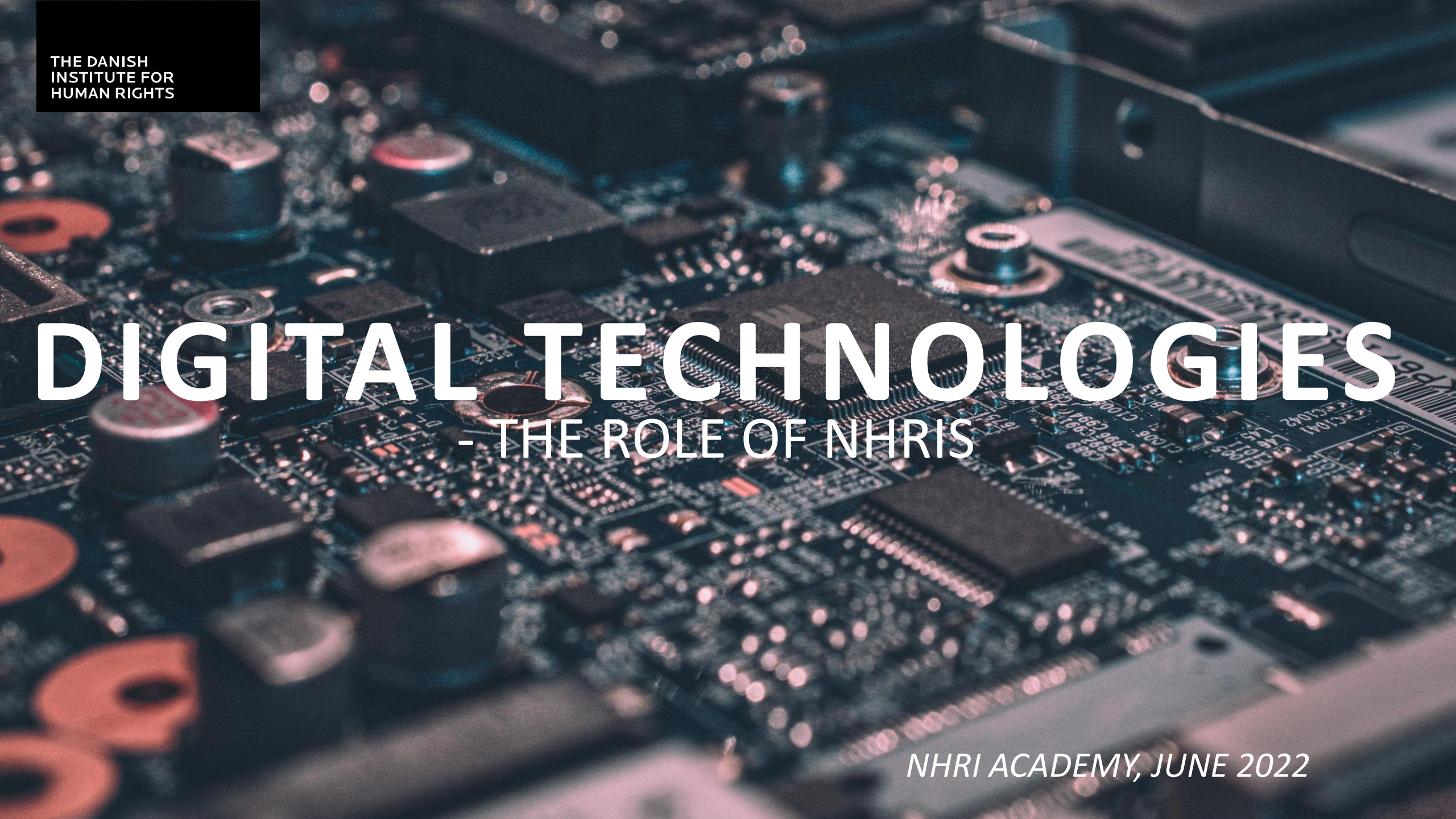


THE DANISH INSTITUTE FOR HUMAN RIGHTS

DIGITAL TECHNOLOGIES

- THE ROLE OF NHRIS

NHRI ACADEMY, JUNE 2022

THE DANISH INSTITUTE FOR HUMAN RIGHTS

Agenda

- E-governance and human rights in tech (Line)
- NHRIs protection and promotion functions (Jane)
- *Group work: how does your NHRI interpret the protection and promotion functions?*
- Application of AI profiling models in public administration: human rights impacts and areas for NHRIs to focus (Jane)
- *Group work: what are the real and potential human rights risks in the nexus of tech and human rights in your country?*
- Examples of tech-related work from other NHRIs (Portugal and the Netherlands)
- *Group work: How would your NHRI address the identified challenges?*

Technology Impacting Human Rights

- Surveillance & social control
- Digital exclusion
- Datafication
- AI & algorithmic profiling
- Leveraging tech for positive impacts

“

”Human rights apply online as off-line”

UN resolution 20/8

PROBLEM:

Governance, policy & enforcement challenges

(part of) SOLUTION:

Safe guarding negative impact of technology through:

- Leveraging human rights frameworks (incl. UNGP)
- Good Governance principles compliance

DIGITAL ETHICS

Responsible AI

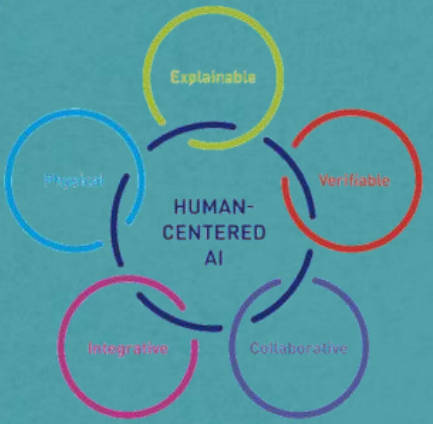
HUMAN RIGHTS BY DESIGN

SUSTAINABLE AI

AI *for* Sustainability and the Sustainability *of* AI.

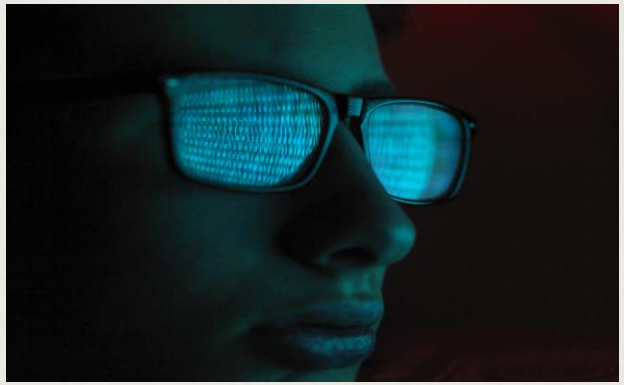
Data Ethics

AI for Good



#Tech4Good
people before software

INTERPRETATION OF THE NHRI PROTECTION AND PROMOTION FUNCTIONS IN RELATION TO TECH



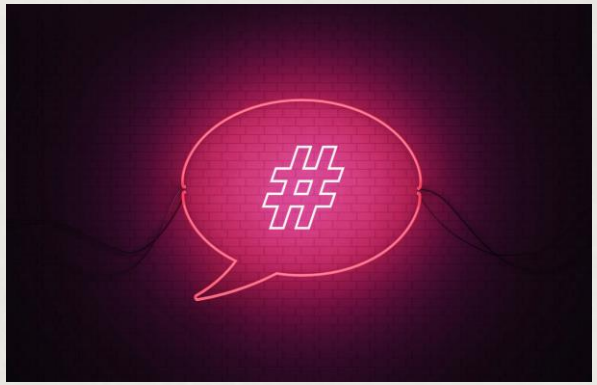
‘Protection’ functions may be understood as those that address and seek to prevent actual human rights violations. Such functions include:

- Human rights investigations
- Investigating individual complaints
- Alternative dispute resolution (mediation and conciliation)
- Inquires (desk audits; informal hearing; formal public hearing)

Example of interpretation: prevent violation of **the minimum** human rights protection required by law (the Danish Institute for Human Rights)

Tech-relevant considerations:

- the issue of jurisdiction
- lack of clear standards
- lack of awareness, hence ”signals” from the rights-holders
- failure to comply is not synonymous with human rights violation
- inadequate resources to process increased number of complaints



‘Promotion’ functions seek to create a society where human rights are more broadly understood and respected. Such functions may include:

- Human Rights Education and Training
- Public Awareness (campaigns, events, thematic seminars, workshops etc.)
- Media strategies (incl. digital media)
- Publications
- Community-based initiatives (documentation centers, cultural events, human rights awards etc.)

Example of interpretation: strengthen commitment to human rights, rather than knowledge and awareness raising; focus on formal education system and TOT rather than direct training (the Danish Institute for Human Rights)

Tech-relevant considerations:

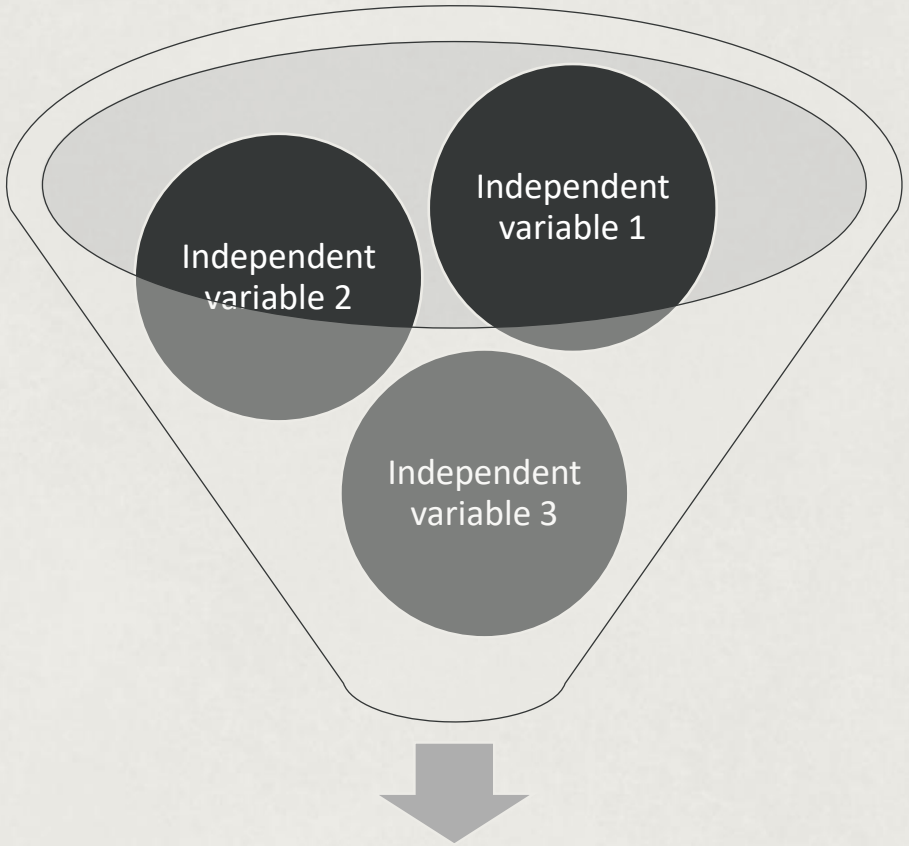
- often overlooked function due to lack of resources
- lack of awareness among duty-bearers and rights-holders
- large amount of actors and stakeholders
- the lack of harmonized and broadly-accepted standards

AI profiling models - How does it all work

Dataset: a number of independent variables (e.g. previous cases; court decisions; statistics etc.) that the model uses to train itself to determine dependent variable

Variables: criteria in the dataset, which the model is trained to identify and compare (e.g. age, gender, income, criminal record etc.)

Weights: each independent variable could be assigned a mathematically defined weight to adjust the model to be as precise as possible



Different types of profiling ai models

AI Model #1: Linear Regression

AI Model #2: Deep Neural Networks

AI Model #3: Logistic Regression

AI Model #4: Decision Trees

AI Model #5: Linear Discriminant Analysis

AI Model #6: Naive Bayes

AI Model #7: Support Vector Machines

AI Model #8: Learning Vector Quantization

AI Model #9: K-nearest Neighbors

AI Model #10: Random Forest

Regression

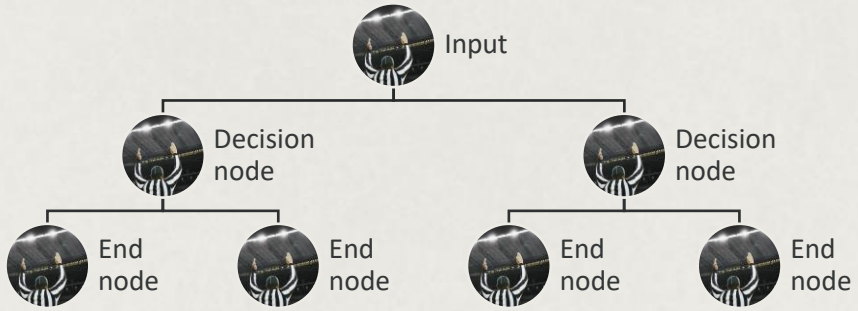
- Determines levels on which the objects possesses dependent variable
- Used to determine e.g. relationship between age, type of illness and lifetime

Classification

- Determines whether the objects possesses dependent variable or not
- Used to determine whether a citizen is eligible for social payment, insurance compensation etc.

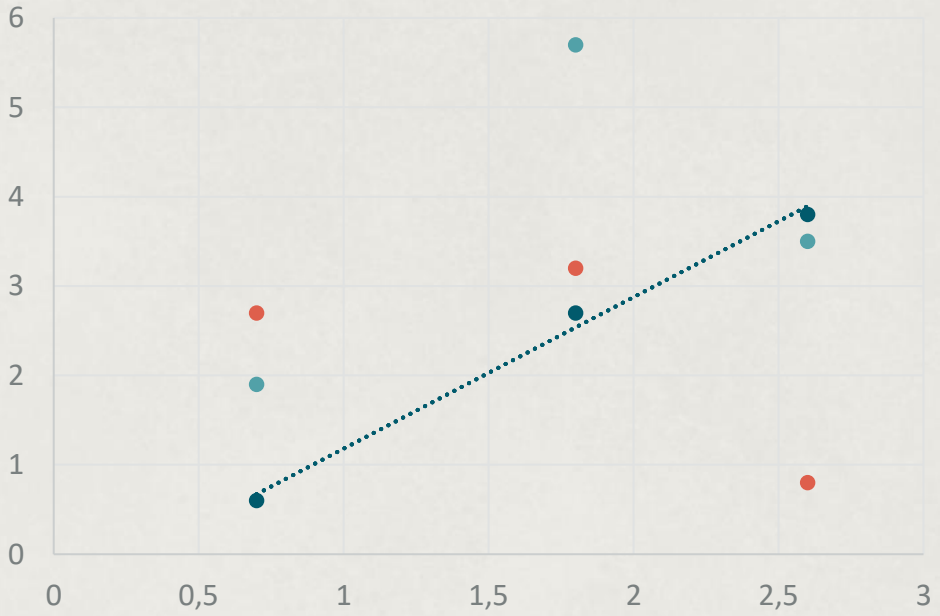
Predictive Decision Trees: decision nodes define criteria that the models uses to split objects into groups; end nodes define criteria for the model to stop and show the end groups (as positive or negative towards the dependent variable).

NB: Includes logistic regression (predicts risks for certain behavior, e.g. criminal regression)



Regression models: most common is linear regression; places objects towards the independent variable based on the criteria introduced through the training set (e.g. determined cases in social care or house prices on the market)

In regression models, the key factor is how the variables are weighted towards the end result and towards one another. The higher the complexity of tab-function (regulates weights), the less transparent the model becomes



WHEN STATE USES AI PROFILING MODELS TO MAKE DECISIONS – KEY FOCUS POINTS FOR NHRIS

Decision to use and design

- Impact assessments are key in deciding whether the model should be used and how
- Training-sets must be solid and representative; otherwise, they transfer bias (garbage in – garbage out)
- Too detailed datasets increase complexity, hence ”black box effect” (and remember intersectionality)

Input

- The case-handling procedure must be open and public with maximum involvement of the citizen
- Data used must be protected (GDPR: data-minimalization and purpose-proportion)

Output – use of end result

- The data must be used according to law and not arbitrary criteria (e.g. political views etc.)
- Decisions with too broad interpretation margin must not be fully automatized (Ombudsman Institutions should define the interpretation margin towards automatization)
- Human-in-the loop (human involved in all decisions); Human-on-the loop (human involved in design and control); human-in-command (human oversight)

Transparency

- **Algorithmics transparency**: access to the works of the model, e.g. how different variables are weighted
- Different levels of transparency: from modular (parts of the model) to wholistic (entire model); from global (societies) to local (individual)
- **Systemic transparency**: access to how the end-results are used by the authorities
- Four major clusters: training of the model; quality and risk assessment; implementation; oversight mechanisms
- Public access and awareness need to be in place when AI models are used to case handling and other state functions (e.g. public database of all models and their functions)

Discrimination

Direct discrimination: when independent variable is used to categories the objects towards independent variable (e.g.gender is weighted in the hiring process). Can be mitigated by desperate learning processes (detaching the protected variables from being used)

Indirect discrimination: when the model reproduces the bias or inequality in the society (e.g. decisions on child custody will reproduce the pattern of gender distribution of custody even if the model is blind to gender-variable)

Negative feedback loop: the model can enhance discrimination by reproducing the bias that created the problem to begin with

Catch 22: the more nuanced the data-set; the more inaccurate the result

NHRI reflection tool on the nexus between human rights and technology

Interactive tool that can be used as a ressource for more knowledge; inspiration and best practice AND a workshop tool for self-assessment; internal planning and/or meetings with external stakeholders

Key target group: NHRIs interested in working with human rights impacts of technology.

Accessible online (through the websites of alliance members?) - material is downloadable for offline use

WHY SHOULD NHRIs ENGAGE WITH THE TECH AND HUMAN RIGHTS NEXUS

This section explains the significance of the human rights and tech-agenda and vice versa. The section includes videos on the tech and human rights nexus, produced by NHRIs and other national and international human rights organisations

MANDATE AREAS

This section describes how NHRIs can address the tech and human rights nexus within their mandate areas. The section is structured through a number of ”cards”. Each card represents a mandate area and is a downloadable PDF (so that it can be used to offline activities). Each card describes:

- Mandate area and its relation to tech
- Human rights risks and opportunities related to the mandate area
- Examples of actions/best practices

ACTORS/STAKEHOLDERS

This section describes how NHRIs can engage with different types of actors when working with tech and human rights. The section is structured through a number of ”cards”. Each card represents an actor and is a downloadable PDF (so that it can be used to offline activities). Each card describes:

- Actor
- Role and obligation
- Types of possible engagement
- Examples of actions/best practices

WORKSHOP GUIDE

This section is a guide for using the cards above for a workshop. The guide includes step-by-step description of possible workshop activities (from planning to evaluation), as well as methodological and didactical considerations. The chapters of the guide can be read online or downloaded as PDFs. The workshops can include:

- Internal workshop within the NHRI
- External workshop for stakeholders
- Partner workshop

NHRI Digital Rights Alliance

- An informal group of NHRIs interested in working towards strengthened human rights compliance of digitalisation
- Purpose: Awareness-raising, capacity-building, peer-learning and joint efforts to strengthen national implementation
- Focus on the NHRI mandate areas as the organising principle
- Activities 2022:
  - Biannual meetings
  - Regular thematic meetings e.g at RightsCon
  - Co-creation of a NHRI Reflection Tool: roles & actors within different mandate areas and best practices
- Contact and more info:
  - Line Gamrath Rasmussen: lira@humanrights.dk