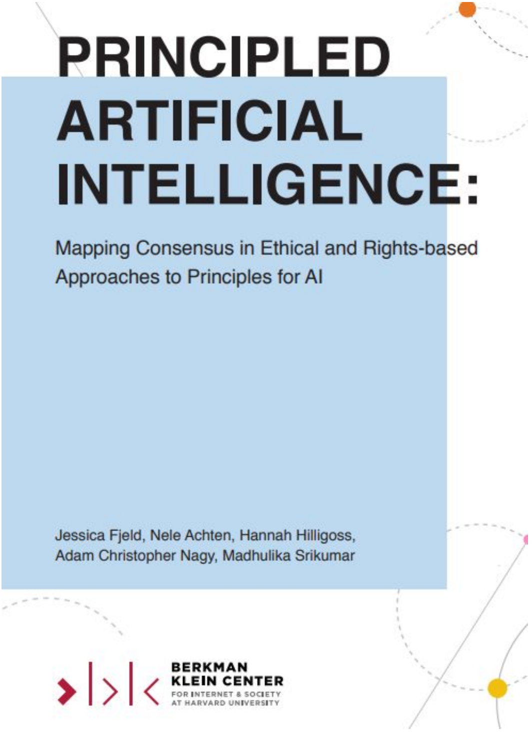
Responsible Guidelines for AA Tasks in NLP

(Work in progress / Preprint by Vageesh Saxena, Aurelia Tamò-Larrieux, Gerasimos Spanakis, Gijs van Dijck)

Responsible AI Frameworks as a Foundation

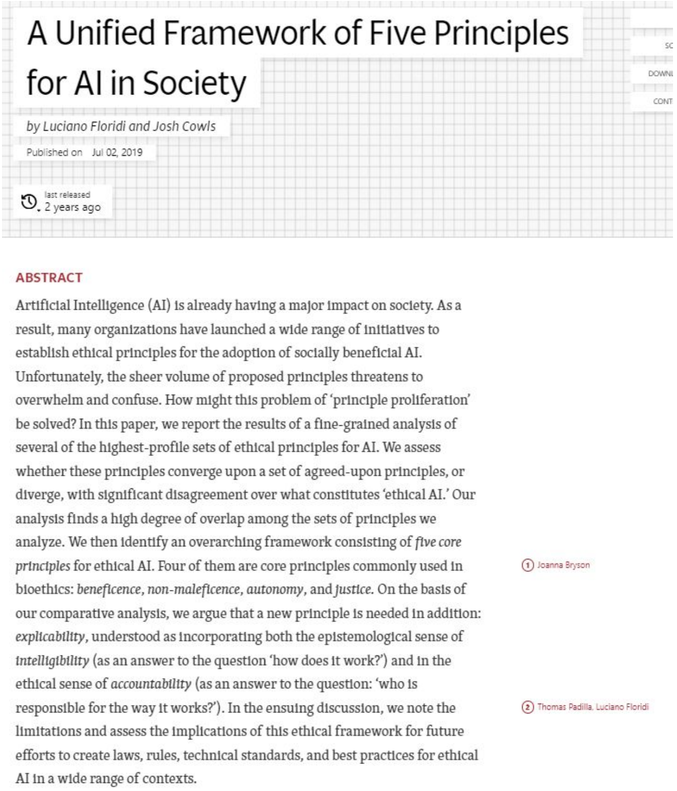


PRINCIPLED ARTIFICIAL INTELLIGENCE:

Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI

Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Christopher Nagy, Madhulika Srikumar

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY



A Unified Framework of Five Principles for AI in Society

by Luciano Floridi and Josh Cowls

Published on Jul 02, 2019

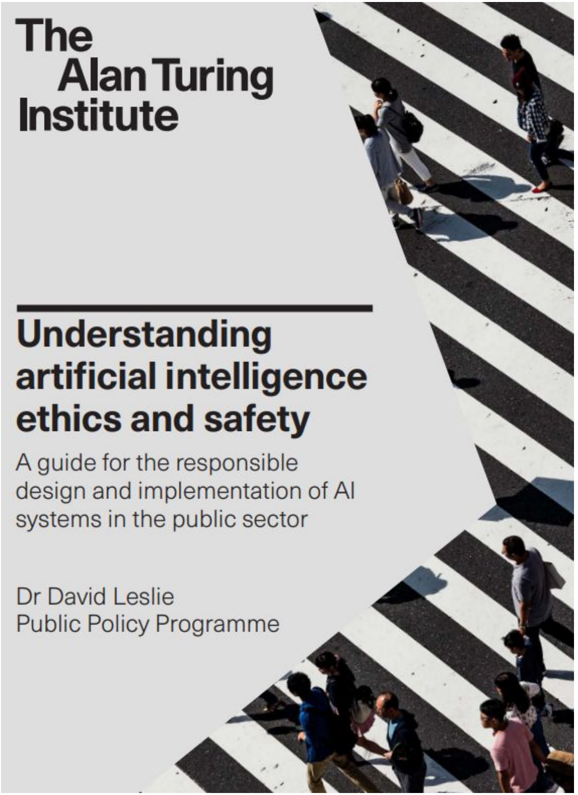
last released
2 years ago

ABSTRACT

Artificial Intelligence (AI) is already having a major impact on society. As a result, many organizations have launched a wide range of initiatives to establish ethical principles for the adoption of socially beneficial AI. Unfortunately, the sheer volume of proposed principles threatens to overwhelm and confuse. How might this problem of 'principle proliferation' be solved? In this paper, we report the results of a fine-grained analysis of several of the highest-profile sets of ethical principles for AI. We assess whether these principles converge upon a set of agreed-upon principles, or diverge, with significant disagreement over what constitutes 'ethical AI.' Our analysis finds a high degree of overlap among the sets of principles we analyze. We then identify an overarching framework consisting of *five core principles* for ethical AI. Four of them are core principles commonly used in bioethics: *beneficence*, *non-maleficence*, *autonomy*, and *justice*. On the basis of our comparative analysis, we argue that a new principle is needed in addition: *explicability*, understood as incorporating both the epistemological sense of *intelligibility* (as an answer to the question 'how does it work?') and in the ethical sense of *accountability* (as an answer to the question: 'who is responsible for the way it works?'). In the ensuing discussion, we note the limitations and assess the implications of this ethical framework for future efforts to create laws, rules, technical standards, and best practices for ethical AI in a wide range of contexts.

(1) Joanna Bryson

(2) Thomas Padilla, Luciano Floridi



The Alan Turing Institute

Understanding artificial intelligence ethics and safety

A guide for the responsible design and implementation of AI systems in the public sector

Dr David Leslie
Public Policy Programme

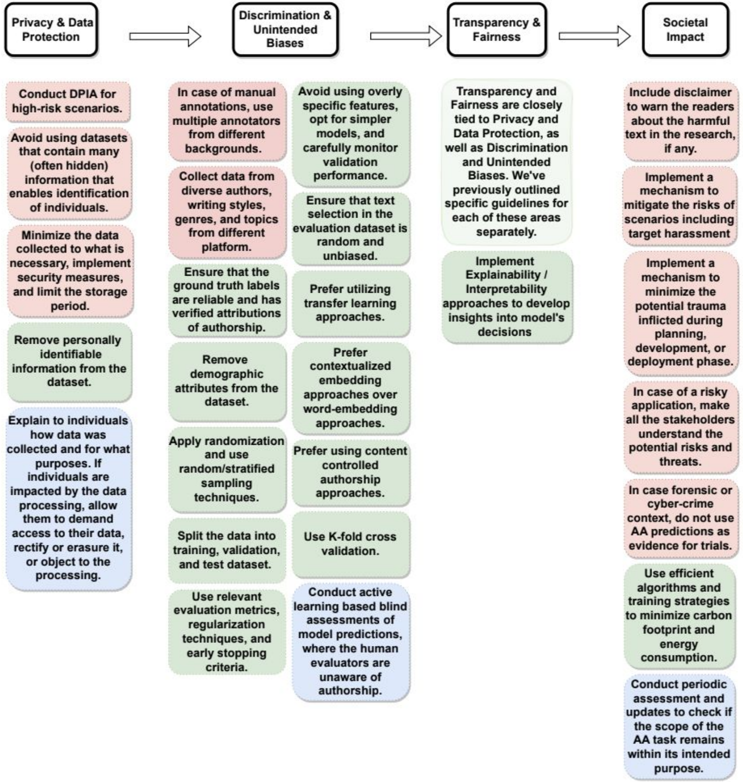
Some Challenges of AI Frameworks

Which one is the most suitable for AA tasks in NLP when dealing with darknet markets? → sensitivity of the domain

How can guidelines be best operationalized? → research shows a lack of adoption (Prem, 2023)

How can we conceptualize the needed steps in different phases and take different stakeholders into account?

Our Approach



Privacy & Data Protection → Discrimination & Unintended Biases → Transparency & Fairness → Societal Impact

Privacy & Data Protection

Conduct DPIA for high-risk scenarios.

Avoid using datasets that contain many (often hidden) information that enables identification of individuals.

Minimize the data collected to what is necessary, implement security measures, and limit the storage period.

Remove personally identifiable information from the dataset.

Explain to individuals how data was collected and for what purposes. If individuals are impacted by the data processing, allow them to demand access to their data, rectify or erasure it, or object to the processing.

Discrimination & Unintended Biases

In case of manual annotations, use multiple annotators from different backgrounds.

Collect data from diverse authors, writing styles, genres, and topics from different platform.

Ensure that the ground truth labels are reliable and has verified attributions of authorship.

Remove demographic attributes from the dataset.

Apply randomization and use random/stratified sampling techniques.

Split the data into training, validation, and test dataset.

Use relevant evaluation metrics, regularization techniques, and early stopping criteria.

Avoid using overly specific features, opt for simpler models, and carefully monitor validation performance.

Ensure that text selection in the evaluation dataset is random and unbiased.

Prefer utilizing transfer learning approaches.

Prefer contextualized embedding approaches over word-embedding approaches.

Prefer using content controlled authorship approaches.

Use K-fold cross validation.

Conduct active learning based blind assessments of model predictions, where the human evaluators are unaware of authorship.

Transparency & Fairness

Transparency and Fairness are closely tied to Privacy and Data Protection, as well as Discrimination and Unintended Biases. We've previously outlined specific guidelines for each of these areas separately.

Implement Explainability / Interpretability approaches to develop insights into model's decisions

Societal Impact

Include disclaimer to warn the readers about the harmful text in the research, if any.

Implement a mechanism to mitigate the risks of scenarios including target harassment

Implement a mechanism to minimize the potential trauma inflicted during planning, development, or deployment phase.

In case of a risky application, make all the stakeholders understand the potential risks and threats.

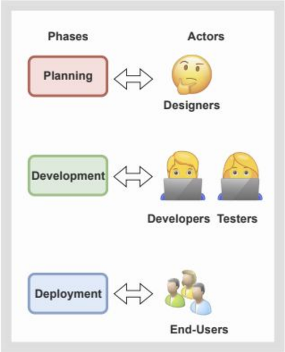
In case forensic or cyber-crime context, do not use AA predictions as evidence for trials.

Use efficient algorithms and training strategies to minimize carbon footprint and energy consumption.

Conduct periodic assessment and updates to check if the scope of the AA task remains within its intended purpose.

Phases

Software Development Life Cycle (SDLC)



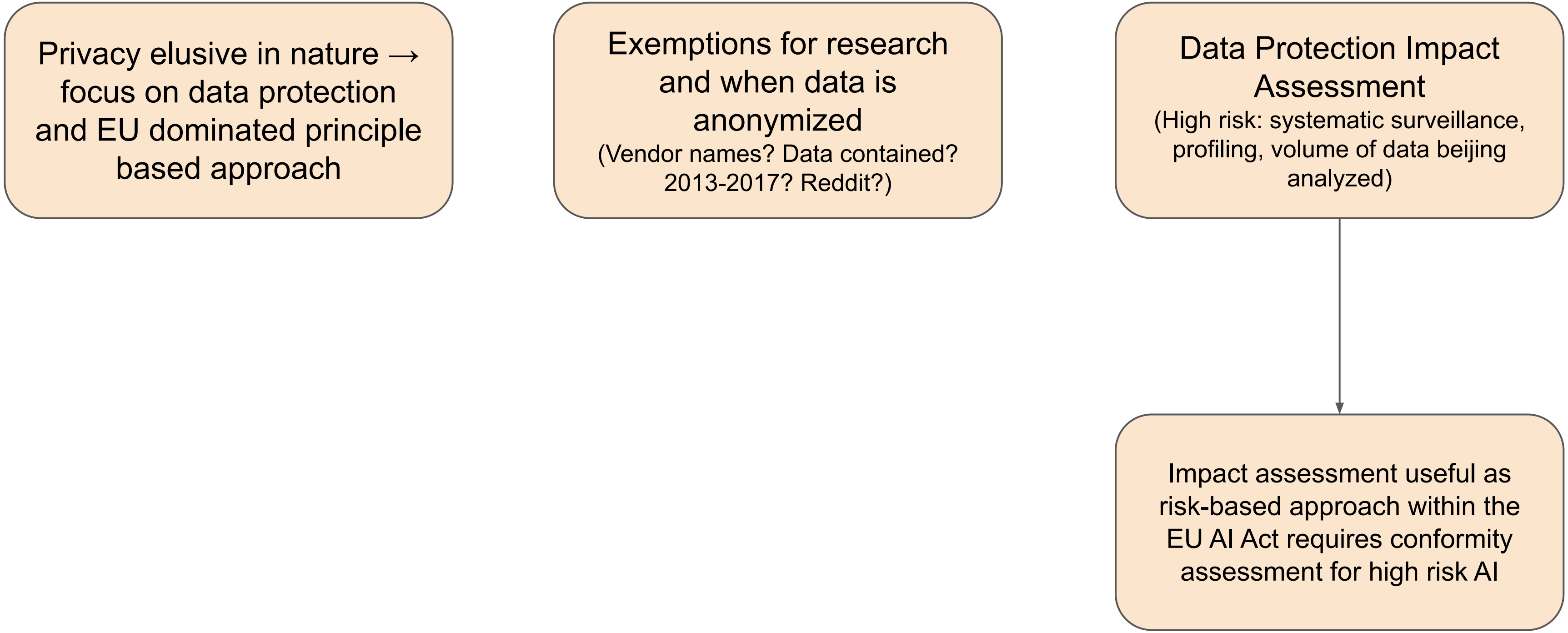
Actors

Phases and actors interplay: Guidelines targeted to the actor best suited to detect and address certain issues

Themes

Relying on established elements that responsible AA guidelines must integrate and adapting them depending on the AA tasks at hand (e.g., VendorLink, TrafficID)

Privacy and data protection



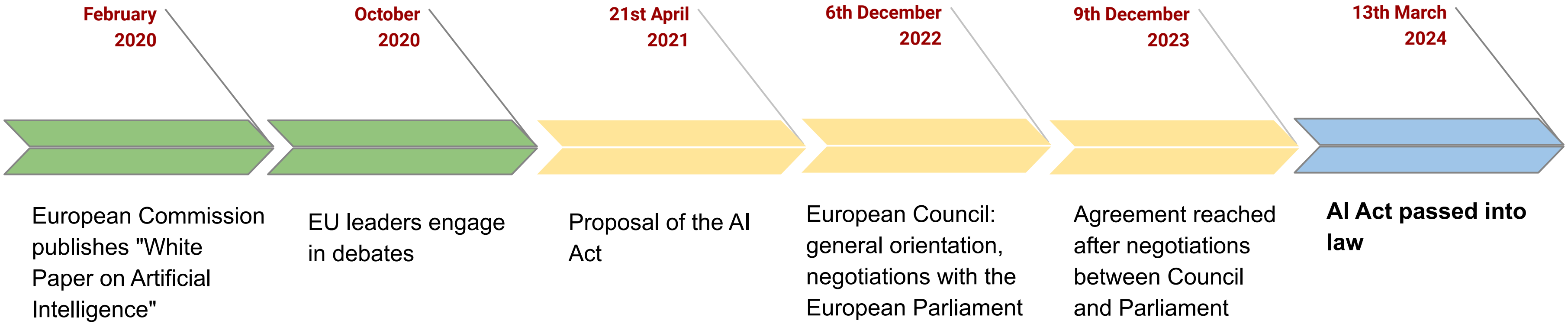
Privacy elusive in nature → focus on data protection and EU dominated principle based approach

Exemptions for research and when data is anonymized
(Vendor names? Data contained? 2013-2017? Reddit?)

Data Protection Impact Assessment
(High risk: systematic surveillance, profiling, volume of data beijing analyzed)

Impact assessment useful as risk-based approach within the EU AI Act requires conformity assessment for high risk AI

AI Act Regulatory History



February 2020

European Commission publishes "White Paper on Artificial Intelligence"

October 2020

EU leaders engage in debates

21st April 2021

Proposal of the AI Act

6th December 2022

European Council: general orientation, negotiations with the European Parliament

9th December 2023

Agreement reached after negotiations between Council and Parliament

13th March 2024

AI Act passed into law

AA in NLP for law enforcement under AI Act

High-Risk AI Systems in Law Enforcement Article 6 (2) and Annex III, section 6 (e):
Law enforcement, ***in so far as their use is permitted*** under relevant Union or national law. AI systems intended to be used by ***or on behalf of law enforcement authorities*** or by Union institutions, bodies, offices or agencies in support of law enforcement authorities **for the profiling of natural persons** as referred to in Article 3(4) of Directive (EU) 2016/680 in the **course of the detection, investigation** or prosecution of criminal offences.

Requirements for providers of high-risk AI systems (Art. 8-25):

1	Risk Management	✗	✓
2	Data Governance	✗	✓
3	Technical documentation	✗	✓
4	Record Keeping	✗	✓
5	Instructions for use	✗	✓
6	Human oversight	✗	✓
7	Accuracy, Robustness and Cybersecurity	✗	✓
8	Quality Management system	✗	✓

Discrimination and unintended bias

Evaluation Bias: incorrect conclusions of the models's capabilities (e.g., words that are repeated a lot may skew the model) → determine evaluation metrics suited for the context

User-Interaction Bias: annotators' feedback can influence the model's training and evaluation → blind assessments, involving diverse users base

Domain and Genre Bias: Different language use depending domains and genres can lead to incorrect AA → balanced representation of authors from different domains and genres

Overfitting and Underfitting: common techniques to address both well known issues in ML exist

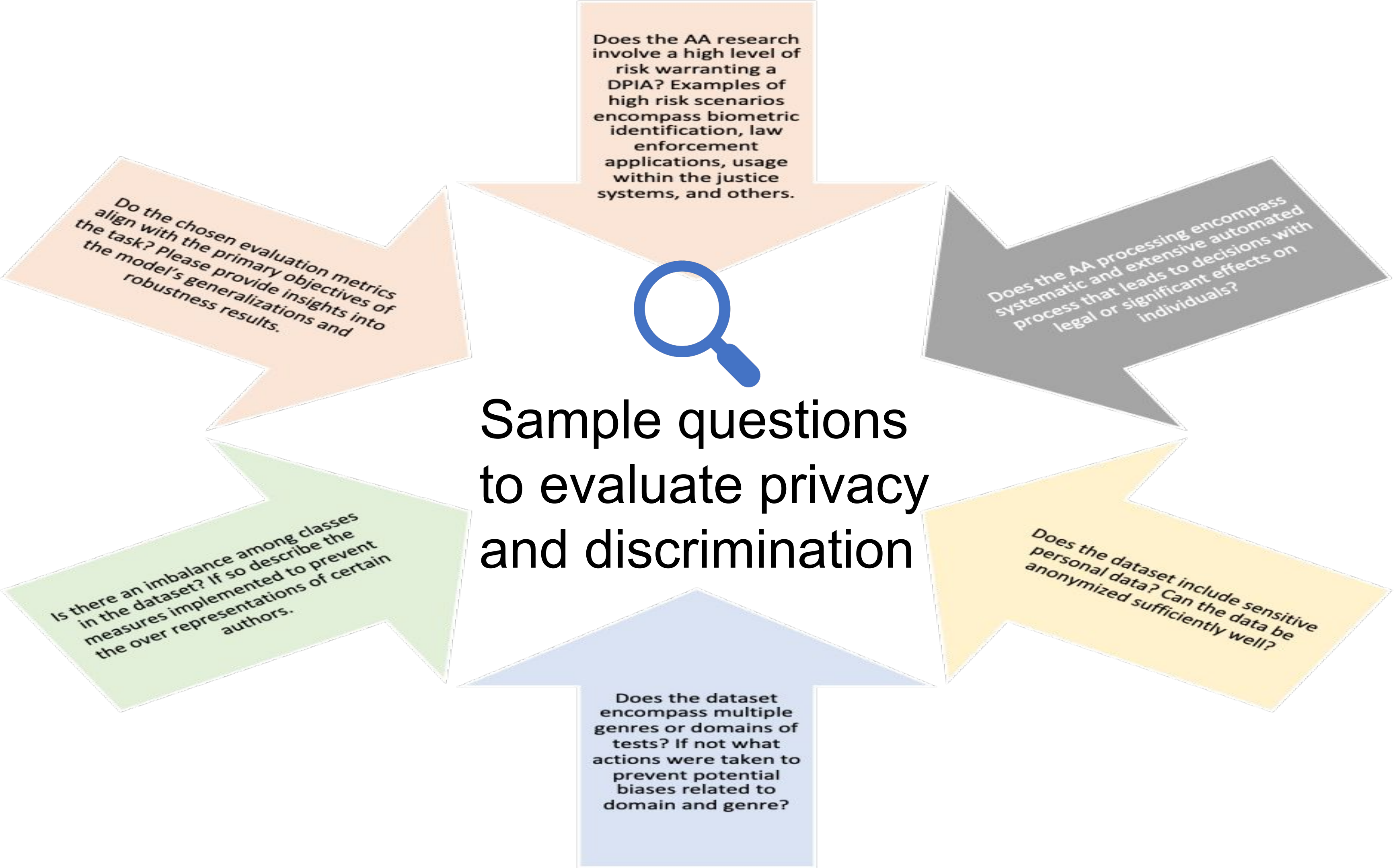
Sampling Bias: inadequate user representation → random and stratified sampling to address this

Label Bias: biased or incorrect attributions within the data → need for verified attributions; active learning strategy to ensure periodic review and update of the training data

Selection Bias: dataset used for training does not accurately represent the target distribution → semi-supervised approaches and combination of data from different platforms

Demographic and Population Bias: correlation with author's demographic attributes can lead to unfair predictions → remove demographic attributes

Sample questions to evaluate privacy and discrimination



Does the AA research involve a high level of risk warranting a DPIA? Examples of high risk scenarios encompass biometric identification, law enforcement applications, usage within the justice systems, and others.

Does the AA processing encompass systematic and extensive automated process that leads to decisions with legal or significant effects on individuals?

Does the dataset include sensitive personal data? Can the data be anonymized sufficiently well?

Does the dataset encompass multiple genres or domains of tests? If not what actions were taken to prevent potential biases related to domain and genre?

Is there an imbalance among classes in the dataset? If so describe the measures implemented to prevent the over representations of certain authors.

Do the chosen evaluation metrics align with the primary objectives of the task? Please provide insights into the model's generalizations and robustness results.

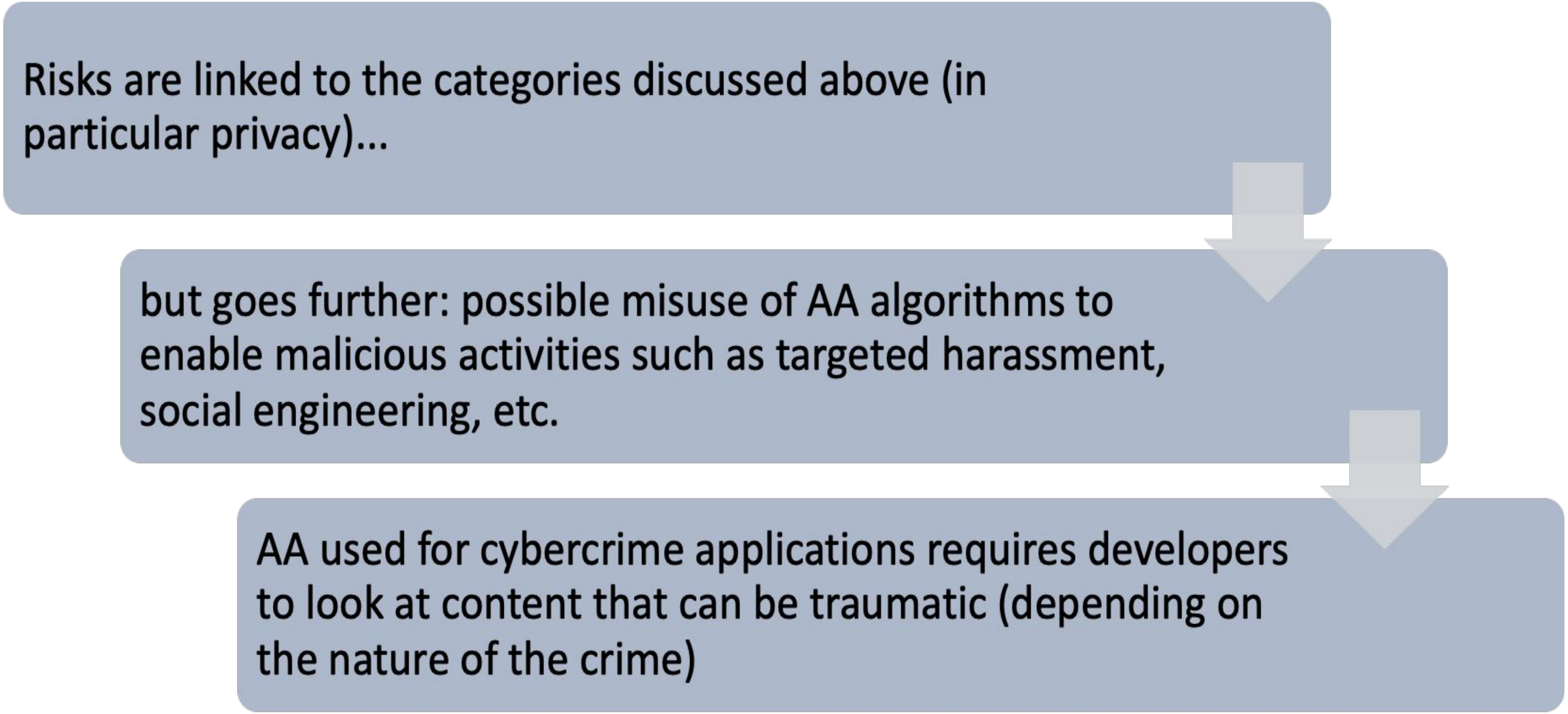
Transparency and fairness

Closely related to privacy and data protection (link over the principles) and discrimination and biases

XAI: Enables detecting also unfair attributions and shedding light on possibly problematic correlations

Challenges of finding the accurate balance between simplicity and and accuracy in explanations

Risks



Risks are linked to the categories discussed above (in particular privacy)…

but goes further: possible misuse of AA algorithms to enable malicious activities such as targeted harassment, social engineering, etc.

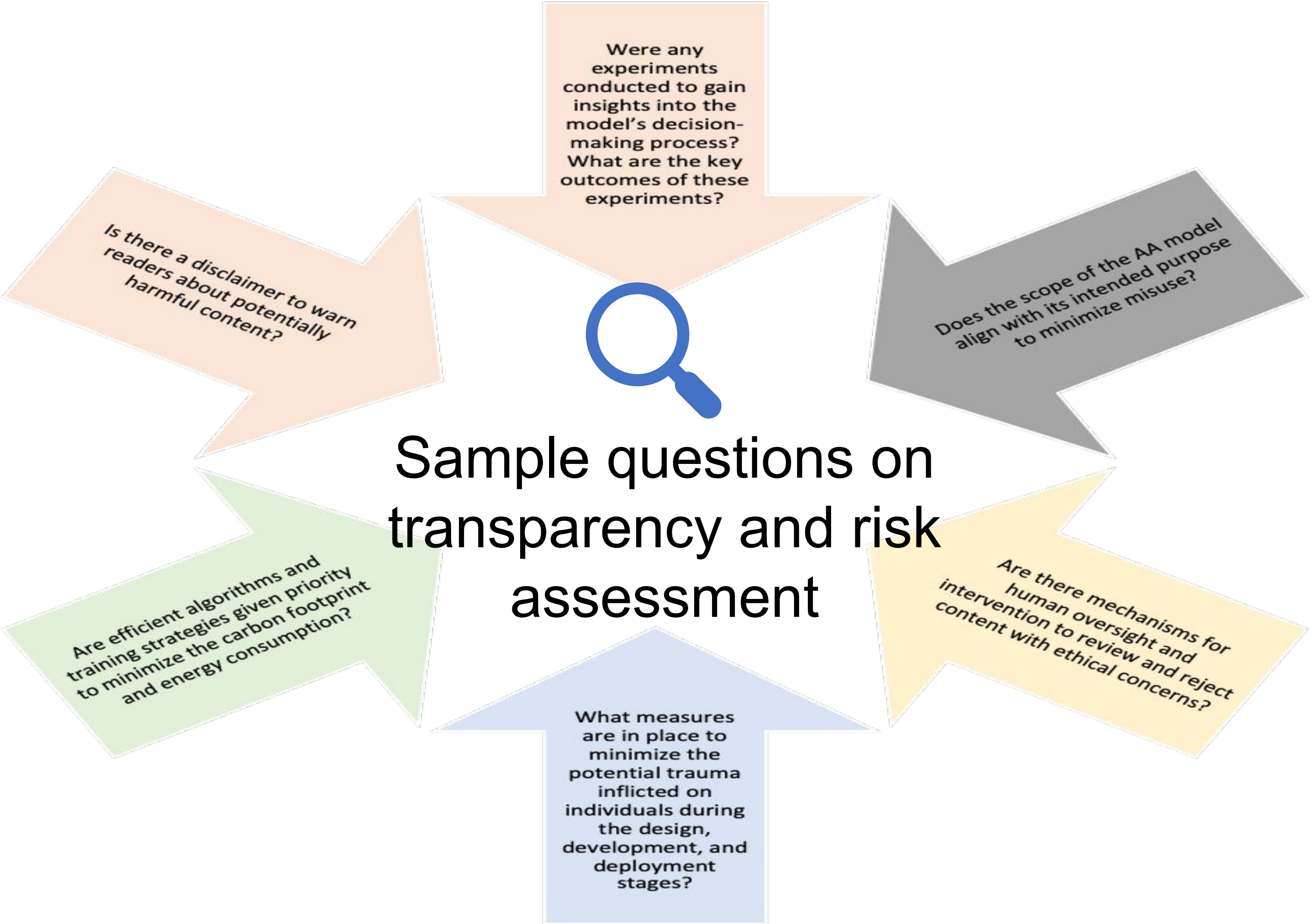
AA used for cybercrime applications requires developers to look at content that can be traumatic (depending on the nature of the crime)

Strategies

strategies to minimize risks to designers, developers, testers, end-users is to map the possible threats and enable collaborative teamwork that enables verbalizing these issues and mental health support

Human in the loop to ensure oversight and itnervation mechanisms and review possible concerns that arise in the development and deployment process of AA algorithms

Minimize the carbon footprint and energy consumption in training AA models → carbon tracking tools



Sample questions on transparency and risk assessment

Were any experiments conducted to gain insights into the model's decision-making process? What are the key outcomes of these experiments?

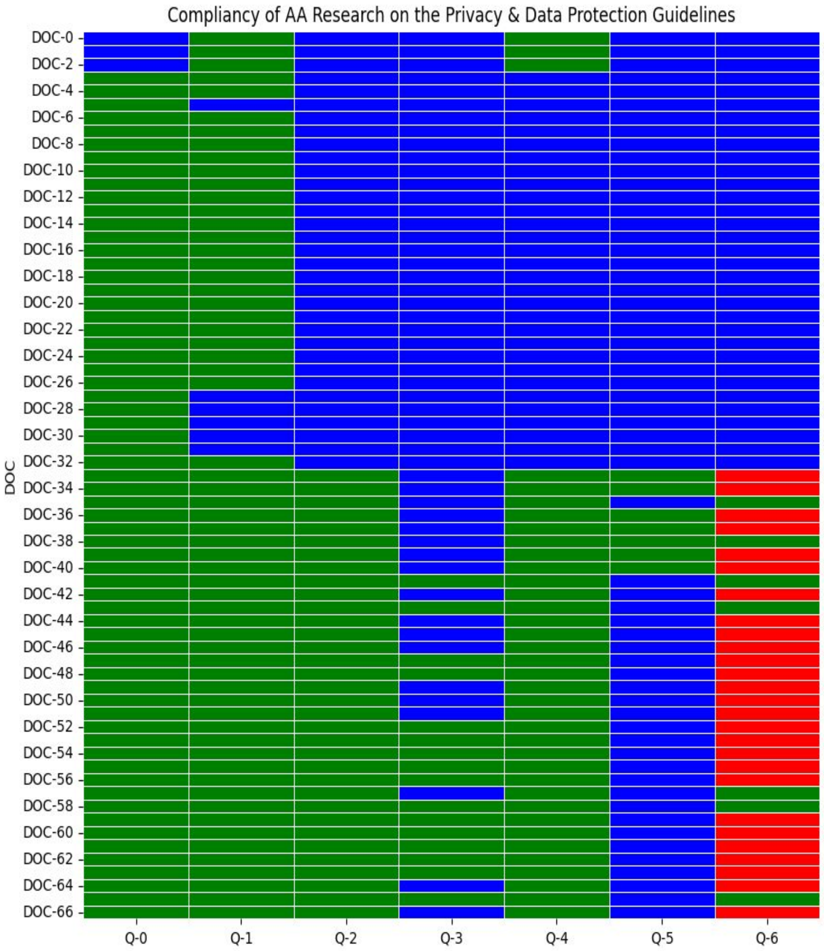
Does the scope of the AA model align with its intended purpose to minimize misuse?

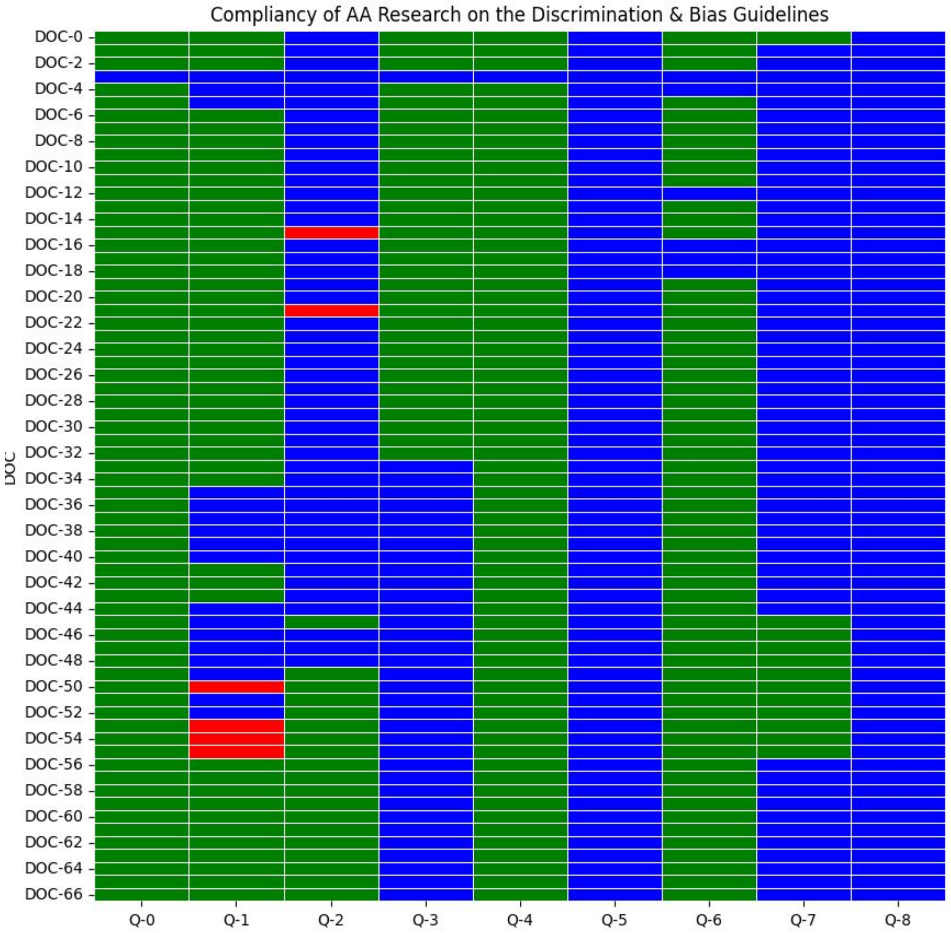
Are there mechanisms for human oversight and intervention to review and reject content with ethical concerns?

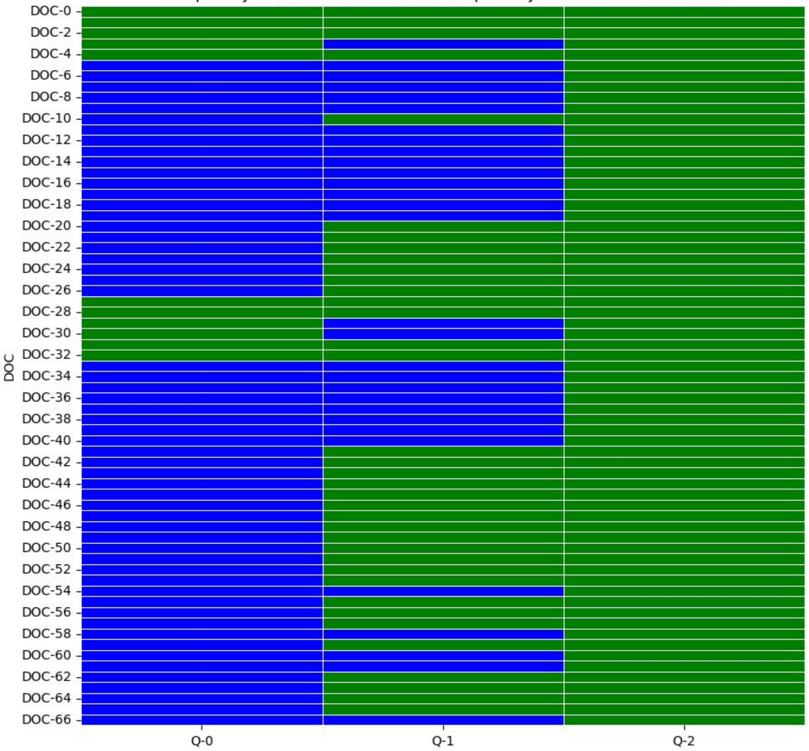
What measures are in place to minimize the potential trauma inflicted on individuals during the design, development, and deployment stages?

Are efficient algorithms and training strategies given priority to minimize the carbon footprint and energy consumption?

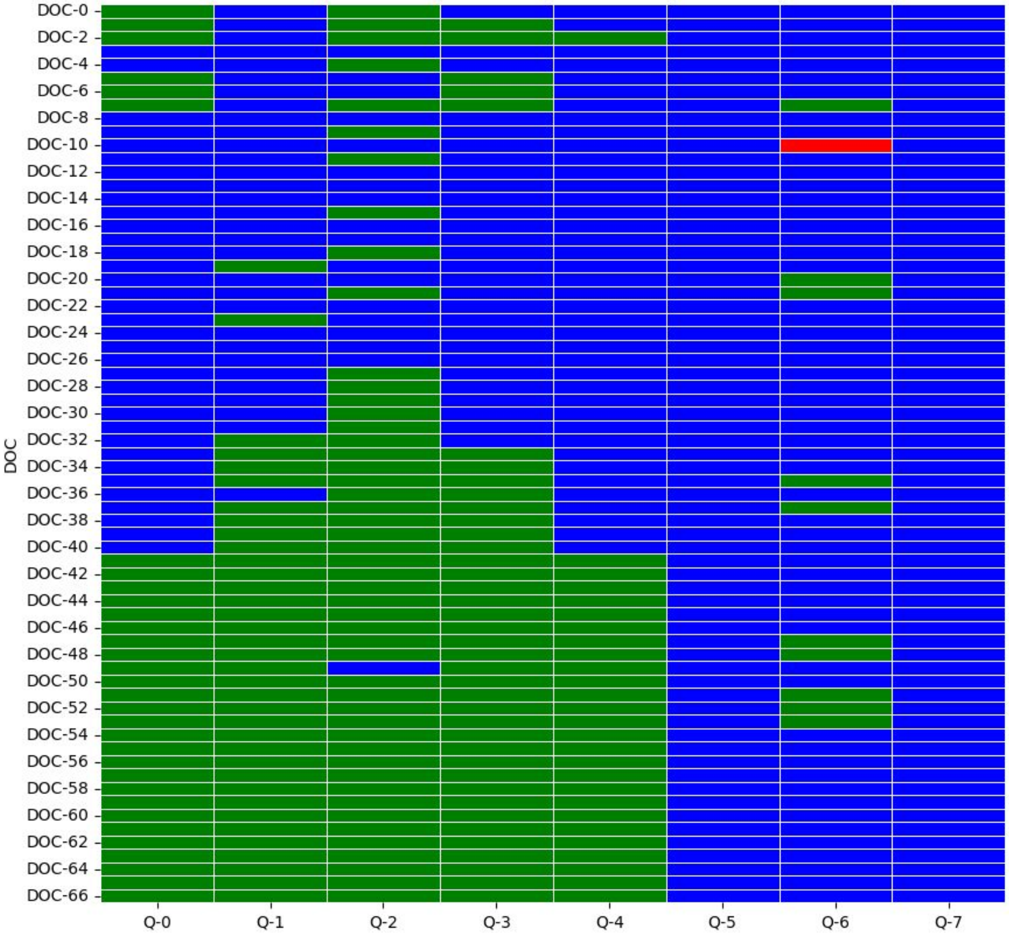
Is there a disclaimer to warn readers about potentially harmful content?


Compliancy of AA Research on the Privacy & Data Protection Guidelines
DOC
DOC-0
DOC-2
DOC-4
DOC-6
DOC-8
DOC-10
DOC-12
DOC-14
DOC-16
DOC-18
DOC-20
DOC-22
DOC-24
DOC-26
DOC-28
DOC-30
DOC-32
DOC-34
DOC-36
DOC-38
DOC-40
DOC-42
DOC-44
DOC-46
DOC-48
DOC-50
DOC-52
DOC-54
DOC-56
DOC-58
DOC-60
DOC-62
DOC-64
DOC-66
Q-0
Q-1
Q-2
Q-3
Q-4
Q-5
Q-6


Compliancy of AA Research on the Discrimination & Bias Guidelines
DOC
DOC-0
DOC-2
DOC-4
DOC-6
DOC-8
DOC-10
DOC-12
DOC-14
DOC-16
DOC-18
DOC-20
DOC-22
DOC-24
DOC-26
DOC-28
DOC-30
DOC-32
DOC-34
DOC-36
DOC-38
DOC-40
DOC-42
DOC-44
DOC-46
DOC-48
DOC-50
DOC-52
DOC-54
DOC-56
DOC-58
DOC-60
DOC-62
DOC-64
DOC-66
Q-0
Q-1
Q-2
Q-3
Q-4
Q-5
Q-6
Q-7
Q-8



Compliancy of AA Research on the Transparency and Fairness Guidelines



Compliancy of AA Research on the Social Impact Guidelines