Call for Input Artificial Intelligence and Human Rights
German Contribution

To date, there is no single **definition of Artificial Intelligence (AI)** accepted by the scientific community. As a result, when dealing with AI a simplified and technologically neutral definition is important, covering those practices or application cases where the development and use of AI systems, or automated decision-making systems more generally, can impact on human rights, democracy and the rule of law, and taking into account all of the systems' socio-technical implications.

Examples for distinct characteristics of AI systems that set them apart from other technologies in relation to their impact on human rights, democracy and the rule of law

- First, the scale, connectedness and reach of AI systems. AI systems analyse an unprecedented amount of fine-grained data (including highly sensitive personal data) at a much faster pace than humans. Moreover, AI systems are subject to statistical error rates. Even if the error rate of a system applied to millions of people is close to zero, thousands of people can still be adversely impacted due to the scale of deployment and interconnectivity of the systems. On the other side, the scale and reach of AI systems also imply that they can be used to mitigate certain risks and biases that are also inherent in other technologies or human behaviour, and to monitor and reduce human error rates.
- Second, the complexity or opacity of many AI systems (in particular in the case of machine learning applications) can make it difficult for humans, including system developers, to understand or trace the system's functioning or outcome. This opacity, may reduce users' trust in AI systems. Moreover, in combination with the involvement of many different actors at different stages during the system's lifecycle, it further complicates the identification of the agent(s) responsible for a potential negative outcome, hence reducing human responsibility and accountability.
- Third, certain AI systems can re-calibrate themselves through feedback and reinforcement learning. However, if an AI system is re-trained on data resulting from its own decisions which contains unjust biases, errors, inaccuracies or other deficiencies, a vicious feedback loop may arise which can lead to a discriminatory, erroneous or malicious functioning of the system and which can be difficult to detect.

AI and Bias / Discrimination

- The scale, connectedness and reach of AI systems can amplify certain risks that are also inherent in other technologies or human behaviour. AI systems analyse an

unprecedented amount of fine-grained data (including highly sensitive personal data) at a much faster pace than humans. This ability can lead AI systems to be used in a way that perpetuates or amplifies unjust bias, also based on new discrimination grounds in case of so called “proxy discrimination”.

- The impact of the use of AI systems on the prohibition of discrimination and the right to equal treatment is one of the most widely reported upon. The use of AI systems can enable the perpetuation and amplification of biases and stereotypes, sexism, racism, ageism and other unjust discriminations, which creates a new challenge to non-discrimination and equal treatment.
- The risk of discrimination can arise for instance due to biased training data (e.g. when the data-set is not sufficiently representative or inaccurate), due to a biased design of the algorithm or its optimisation function (e.g. due to the conscious or unconscious stereotypes or biases of developers), due to exposure to a biased environment once it is being used, or due to a biased use of the AI system.
- For instance, in light of past legal or factual discriminations against women, historical data bases can lack sufficiently gender-balanced data. When such a data base is subsequently used by AI systems, this can lead to equally biased decisions and hence perpetuate unjust discrimination. The same holds true for traditionally vulnerable, excluded or marginalised groups more generally. The gaps in representation of the above-mentioned groups in the AI sector might further amplify this risk.
- In addition, when the transparency of AI systems’ decision-making processes is not ensured, and when mandatory reporting or auditability requirements are not in place, the existence of such biases can easily remain undetected or even be obscured, and thus marginalise the social control mechanisms that typically govern human behavior.
- On the other hand, AI systems can be deployed to detect and mitigate human bias. Also, measures to ensure gender balance in the AI workforce and to improve diversity in terms of ethnic/social origin could help mitigate some of those risks.

Some of **the key obligations when dealing with AI are**:

(a) Human Dignity

- Tasks that risk violating human dignity if carried out by machines rather than human beings, should be reserved for humans.
- AI deployers should inform human beings of the fact that they are interacting with an AI system rather than with a human being whenever confusion may arise

(b) Prevention of harm to human rights, democracy and the rule of law

- States should ensure that developers and deployers of AI systems take adequate measures to minimise any physical or mental harm to individuals, society and the environment. This could, for instance, be done by ensuring that potentially harmful AI systems operate based on an opt-in instead of an opt-out model. Where this is not possible, clear instructions should be provided on how individuals can opt-out from the system's use and on which alternative non-AI driven methods are available.
- States should ensure the existence of adequate (by design) safety, security and robustness requirements and compliance therewith by developers and deployers of AI systems. These requirements should include, inter alia, resilience to attacks, accuracy and reliability, and the necessity to ensure data quality and integrity. Moreover, AI systems should be duly tested and verified prior to their use as well as throughout the entire life cycle of the AI system including by means of periodical reviews to minimise such risks.
- States should ensure that AI systems are developed and used in a sustainable manner, with full respect for applicable environmental protection standards.

(c) Human Freedom and Human Autonomy

- Any AI-enabled manipulation, individualised profiling and predictions involving the processing of personal data must comply with the obligations set out in international law.

States should require AI developers and deployers to establish human oversight mechanisms that safeguard human autonomy, in a manner that is tailored to the specific risks arising from the context in which the AI system is developed and used:

  - An adequate level of human involvement should be ensured in the operation of AI systems, based on a contextual risk assessment taking into account the system's impact on human rights, democracy and the rule of law.
  - Whenever necessary and possible, based on a thorough risk assessment, a qualified human being should be able to disable any AI system or change its functionality.
  - Those developing and operating AI systems should have the adequate competences or qualifications to do so, to ensure appropriate oversight that enables the protection of human rights, democracy and the rule of law.
  - To protect the physical and mental integrity of human beings, AI deployers should strive to avoid the use of 'attention economy' models that can limit human autonomy.

- States should require AI developers and deployers to duly and timely communicate options for redress.

(d) Non-Discrimination, Gender Equality, Fairness and Diversity

- States should ensure that the AI systems they deploy do not result in unlawful discrimination, harmful stereotypes
- States should include non-discrimination and promotion of equality requirements in public procurement processes for AI systems, and ensure that the systems are independently audited for discriminatory effects prior to deployment. AI systems should be duly tested and verified prior to their use as well as throughout the entire life cycle of the AI system including by means of periodical audits and reviews.
- States should impose requirements to effectively counter the potential discriminatory effects of AI systems deployed by both the public and private sectors and protect individuals from the negative consequences thereof. Such requirements should be proportionate to the risks involved, concern the AI systems' entire lifecycle and concern, inter alia, filling existing gender data gaps, the representativeness, quality and accuracy of data sets, the design and optimisation function of algorithms, the use of the system, and adequate testing and evaluation processes to verify and mitigate the risk of discrimination.
-

(e) Principle of Transparency and Explainability of AI systems

- States should require developers and deployers of AI systems to provide adequate communication: Users should be clearly informed of their right to be assisted by a human being whenever using an AI system that can impact their rights or similarly significantly affect them, particularly in the context of public services, and of how to request such assistance.
- States should impose requirements on AI developers and deployers regarding traceability and the provision of information:
    - Persons with a legitimate interest (e.g. consumers, citizens, supervisory authorities or others) should have easy access to contextually relevant information on AI systems.
    - This information should be comprehensible and accessible and could, inter alia, include the types of decisions or situations subject to automated processing, criteria relevant to a decision, information on the data used, a description of the method of the data collection. A description of the system's potential legal or other effects should be accessible for review/audit by independent bodies with necessary competences.
    - Specific attention should be paid if children or other vulnerable groups are subjected to interaction with AI systems.
- States should impose requirements on AI developers and deployers regarding documentation: The data sets and processes that yield the AI system's decisions, including those of data gathering, data labelling and the algorithms used, should be

documented, hence enabling the ex post auditability of the system. Qualitative and effective documentation procedures should be established.

(f) Data protection and the right to privacy

- States should ensure that the right to privacy and data protection are safeguarded throughout the entire lifecycle of AI systems that they deploy, or that are deployed by private actors. The processing of personal data at any stage, including data sets, of an AI system's lifecycle must be based on the principles set out under the Convention 108+.
- States should take particular measures to effectively protect individuals from AI-driven mass surveillance, for instance through remote biometric recognition technology or other AI-enabled tracking technology, if such measures contravene international human rights law.

(g) Accountability and responsibility

- effective remedies must be available under respective national jurisdictions, including for civil and criminal responsibility, and that accessible redress mechanisms are put in place.
- States should establish public oversight mechanisms for AI systems that may breach legal norms in the sphere of human rights, democracy or the rule of law.
- States should ensure that developers and deployers of AI systems identify, document and report on potential negative impacts of AI systems on human rights, democracy and the rule of law and put in place adequate mitigation measures to ensure accountability for any caused harm.
- States should put in place measures to ensure that public authorities are always able to audit AI systems used by private actors, so as to assess their compliance with existing legislation and to hold private actors accountable.

(h) Democracy

- States should take adequate measures to counter the use or misuse of AI systems for unlawful interference in electoral processes, for personalised political targeting without adequate transparency, responsibility and accountability mechanisms, or more generally for shaping voters' political behaviours or to manipulate public opinion in a manner that can breach legal norms safeguarding human rights, democracy and the rule of law.
- States should adopt strategies and put in place measures for fighting disinformation and identifying online hate speech to ensure fair informational plurality.

- States should make public and accessible all relevant information on AI systems (including their functioning, optimisation functioning, underlying logic, type of data used) that are used in the provision of public services, while safeguarding legitimate interests such as public security.
- States should put in place measures to increase digital literacy and skills in all segments of the population. Their educational curricula should adjust to promote a culture of responsible innovations that respects human rights, democracy and the rule of law.

(i) Rule of Law

- States must ensure that AI systems used in the field of justice and law enforcement are in line with the essential requirements of the right to a fair trial. They should pay due regard to the need to ensure the quality and security of judicial decisions and data, as well as the transparency, impartiality and fairness of data processing methods. Safeguards for the accessibility and explainability of data processing methods, including the possibility of external audits, should be introduced to this end.
- States must ensure that effective remedies are available and that accessible redress mechanisms are put in place for individuals whose rights are violated through the development or use of AI systems in contexts relevant to the rule of law.
- States should provide meaningful information to individuals on the use of AI systems in the public sector whenever this can significantly impact individuals' lives. Such information must especially be provided when AI systems are used in the field of justice and law enforcement, both as concerns the role of AI systems within the process, and the right to challenge the decisions informed or made thereby.
- States should ensure that use of AI systems does not interfere with the decision-making power of judges or judicial independence and that any judicial decision is submitted to human oversight.

(4) Compliance mechanisms

- Practical compliance mechanisms (such as impact assessments, lifecycle auditing, and monitoring, certification methods, and sandboxes) are one way of driving such compliance and of helping States to understand and monitor adherence to the legal framework.

- Such mechanisms confer further benefits beyond compliance, for example by increasing transparency around the use of AI and creating a common framework for promoting trust.
- Compliance mechanisms might be used to assess the design of an AI-enabled system, as well as its operational processes, contextual implementation and use case. On the question of when AI systems that have an impact on human rights, democracy and the rule of law should be subject to such assessment, ex ante assessment and continuous assessment at various milestones throughout the AI project lifecycle, including after initial deployment and use, are important. Compliance mechanisms should also evolve over time to account for the evolving nature of the system. To ensure that impact assessments can be used efficiently, particular attention should be paid to their comprehensibility and accessibility to all relevant actors. Legal safeguards should ensure that compliance mechanisms are not used by organisations to shield themselves from potential liability claims associated with their conduct.
- Examples:
    - Human Rights Impact Assessment
    - Certification and Quality Labelling
    - Audits
    - Regulatory Sandboxes
    - Continuous Monitoring