unesco

GLOBAL TOOLKIT ON

AI AND THE RULE OF LAW FOR THE JUDICIARY

Published in 2026 by the United Nations Educational, Scientific and Cultural Organization,
7, place de Fontenoy, 75352 Paris 07 SP, France
© UNESCO 2026
ISBN: 978-92-3-100879-5
DOI: https://doi.org/10.58338/DNRE2540

This publication is available in Open Access under the Attribution-ShareAlike 3.0 IGO (CC-BY-SA 3.0 IGO) license (https://creativecommons.org/licenses/by-sa/3.0/igo/deed.en). By using the content of this publication, the users accept to be bound by the terms of use of the UNESCO Open Access Repository (https://www.unesco.org/en/open-access/cc-sa).

Images marked with an asterisk (*) do not fall under the CC-BY-SA license and may not be used or reproduced without the prior permission of the copyright holders.

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

This toolkit was prepared by:

Dr. Miriam Stankovich, Principal Digital Policy Specialist at the Center for Digital Acceleration (Bethesda, Maryland, United States). The section on AI bias and gender equality was developed by Ivana Feldfeber (Co-Founder & Executive Directress of DataGénero), Yasmín Quiroga (Co-Founder of DataGénero & Secretary at Criminal Court N°10 of Buenos Aires, Argentina), and Marianela Ciolfi Felice (Assistant Professor in Interaction Design at KTH University, Sweden, & Advisor at DataGénero). The section on Opportunities: AI and the Judiciary on the African Continent was written by Prof. Vukosi Marivate (University of Pretoria, South Africa).

Academic Advisors:

Prof. Joan Barata Mir (Senior Legal Fellow at Justitia, Denmark-United States), Prof. Maria Fasli (University of Essex, United Kingdom), Prof. Els de Busser (Leiden University, Netherlands), and Prof. Vukosi Marivate (University of Pretoria, South Africa).

UNESCO Reviewers:

Cedric Wachholz, Jaco Du Toit, Bhanu Neupane, Rosa María González, Natalia Zuazo, Misako Ito, Mehdi Benchelah.

External Reviewers:

Jhalak M. Kakkar (Executive Director, Centre for Communication Governance, National Law University Delhi & Visiting Professor, NLU Delhi), Nidhi Singh (Programme Officer, Centre for Communication Governance, NLU Delhi), Judge Jean Aloise Ndiaye (Supreme Court of Senegal), Dr. Alexandre Barbosa (Head of the Regional Center for Studies on the Development of the Information Society, Cetic.br | NIC.br), Luiz Costa (Brazilian Observatory of Artificial Intelligence, OBIA), Ameen Jauhar (Team Lead, ALTR, Vidhi Centre for Legal Policy), Nathalie Smuha (Assistant Professor at KU Leuven Faculty of Law and Emile Noël Fellow at New York University School of Law), Lee Tiedrich (Distinguished Faculty Fellow, Ethical Tech at Duke University & GPAI and OECD AI expert), Marc Rotenberg (President & Founder of the Center for AI and Digital Policy), Alfonso Peralta Gutiérrez (Judge of the First Instance and Criminal Investigation, Granada, Spain), Murali Sagi (Deputy Chief Executive at Judicial Commission of NSW, New South Wales), Anthony Wong (President of IFIP, International Federation for Information Processing), Saurabh Karn (Founder and Lead Scientist at OpenNyAI & Founder of Sampatti Card), and Prof. Keith R. Fisher (Distinguished Fellow, National Judicial College, US), Niki Iliadis (Director, AI and the Rule of Law at TFS, The Future Society), Amanda Leal (Associate, AI Governance at TFS), Nicolas Miailhe (Founder & President of TFS), Prof. Srikrishna Deva Rao (Vice Chancellor at NALSAR University of Law, Hyderabad), Mr. Pranav Verma (Assistant Professor at National Law School of India University, Bengaluru), Dr. Ravi Srinivas (Adjunct Professor at NALSAR University of Law). Dr. Naveen Thayyil (Associate Professor at IIT, Delhi), Neela Badami (Partner at Samvad Partners), Dr. Shouvik Kumar Guha (Associate Professor at West Bengal National University of Juridical Sciences, Kolkata), Rohan George (Partner at Samvad Partners), Nehaa Chaudhari (Partner at Ikigai Law), Pallavi Sondhi (Senior Associate at Ikigai Law), Ajey Karthik (Associate at Ikigai Law) and Namratha Murugeshan (Associate at Ikigai Law), Jaideep Reddy (Technology Lawyer at Trilegal & Visiting Professor at National Law School of India University, Bengaluru).

Project management and coordination:

Prateek Sibal, Programme Specialist, Digital Policies and Digital Transformation, UNESCO.
Charline d'Oultremont, Consultant, Digital Policies and Digital Transformation, UNESCO.
Giovanni Imperiali, Intern, Digital Policies and Digital Transformation, UNESCO.
Gustavo Fonseca Ribeiro, Consultant, Digital Policies and Digital Transformation, UNESCO, contributed to the organization of pilot workshops for the toolkit.
Dr. Kamel El Hilali, AI & Rule of Law Specialist, Digital Policies and Digital Transformation, UNESCO.

Cover and Graphic design: Nube Consulting
Typeset: Nube Consulting + The Amaj
Illustrations: © Shutterstock*
Printed by: UNESCO

Funded by
the European Union

The Global Toolkit has been developed as part of the European Commission funded project "Supporting Member States in Implementing UNESCO's Recommendation on the Ethics of AI through Innovative Tools"

SHORT SUMMARY

Artificial intelligence as a new frontier for the Judiciary

What is Artificial Intelligence (AI)? How does it work? And more importantly, how does it find its way into the judicial context? Technologies such as AI have been around for decades, but only recently have they begun to be used in a variety of justice and law enforcement settings. While AI has immense potential for the justice system, helping judges make better decisions, improving efficiency, increasing access, and helping to detect and prevent crime, there are also some important issues that justice stakeholders should consider as they prepare for a future in which AI is increasingly used in justice systems.

In 2022, UNESCO launched two needs assessments. First, through UNESCO's Artificial Intelligence Needs Assessment Survey in Africa, 90% of the 32 countries surveyed requested capacity building support for the Judiciary on AI. At the same time, a second global survey of judicial actors in 100 countries underlined the need for better understanding the use of AI in the administration of justice and its wider legal implications on societies.

35 000

Judicial Actors from over 160 countries

Effectively Involved by UNESCO's Judges' Initiative

The "Global Toolkit on AI and the Rule of Law" for the Judiciary responds to these needs and provides judicial actors (judges, prosecutors, state attorneys, public lawyers, law universities and judicial training institutions) with the knowledge and tools necessary to understand the benefits and risks of AI in their work. The toolkit will assist judicial actors in mitigating the potential human rights risks of AI by providing guidance on the relevant international human rights laws, principles, rules and emerging jurisprudence that underpin the ethical use of AI.

unesco

"Since wars begin in the minds of men and women, it is in the minds of men and women that the defences of peace must be constructed".

GLOBAL TOOLKIT ON

AI AND THE RULE OF LAW FOR THE JUDICIARY

FOREWORD

Judges play a crucial role in protecting civil rights: they can set powerful legal precedents in their judgements on individual cases, enabling a country to leap forward in a particular area of the law. Recent legal cases have shown that the Judiciary can draw upon international human rights law, constitutional safeguards and data protection laws to safeguard against discriminatory and biased AI systems. If judges are to play this vital role effectively, we must help build their knowledge and understanding of how AI systems work, and how international human rights law can be applied to AI.

Since 2014 UNESCO's Global Judges Initiative has involved over 34,800 judicial actors from over 160 countries on freedom of expression, access to information, and safety of journalists. This initiative helps strengthen the capacities of judicial operators to engage with emerging challenges for the Judiciary and protect fundamental human rights and freedom of expression.

In 2022, the Judges Initiative launched its programme on AI and the Rule of Law with the aim of engaging stakeholders within justice systems in a global and timely discussion on the applications of artificial intelligence and its impact to the rule of law. This follows up on UNESCO's Recommendation on the Ethics of Artificial Intelligence, a comprehensive blueprint for building regulatory regimes upon universally accepted values and principles, adopted by UNESCO's 193 Member States in November 2021. The Recommendation underlined the value of "AI systems to improve access to information and knowledge" and the need to "enhance the capacity of the Judiciary to make decisions related to AI systems as per the rule of law and in line with international law and standards".

Following a worldwide survey involving judicial actors from the Global Judges Initiative Alumni Network, UNESCO and partners developed a Massive Open Online Course on AI and the Rule of Law (MOOC) in seven languages in 2022. The MOOC unpacks good practices on how courts are deciding AI-related cases, according to human rights and ethical standards, and explores the opportunities and risks of AI adoption by justice systems.

In the footsteps of this MOOC, the "Global Toolkit on AI and the Rule of Law" aims to train judicial actors on how to ensure that the development of AI reaches its full potential in accordance with the rule of law. In fact, while we strive to develop new laws to govern AI itself, it is imperative that we support judges, prosecutors and public servants with enhanced capacities to safeguard us from AI-related risks.

Contents

List of Acronyms **14**

Why this Toolkit? **16**

Glossary **20**

Module 1 — Introduction to AI and the Rule of Law **24**

1. Understanding AI and its building blocks 25
2. Why is data important in the context of AI? 35
3. AI systems as "black boxes" 39
4. The human in the loop principle 43
5. Why is cybersecurity important in the context of AI? 46
6. Activities 49
7. Resources 52

Module 2 — AI Adoption in the Judiciary **54**

1. What are the applications of AI in the Judiciary? 55
2. Case studies on AI deployment in the Judiciary 78
3. Activities 83
4. Resources 86

Module 3 — Legal and Ethical Challenges of AI **88**

1. What is AI Ethics? 89
2. What is AI bias? 94
3. Why algorithmic transparency and accountability are important in the context of the Judiciary? 109
4. Spotlight on biometric identification, facial recognition technology, and deepfakes 113
5. Activities 122
6. Resources 127

Module 4 — Human Rights and AI **128**

1. Introduction to human rights and AI 129
2. Select human rights impacted by AI deployment 136
3. Approaches to AI governance 184
4. Activities 191
5. Resources 194

Suggested UNESCO Resources **198**

How to Use This Toolkit? **201**

Annex I - UNESCO Ethical Impact Assessment for AI systems **204**

Annex II - Examples of additional activities **206**

Annex III — Training agenda — template **209**

List of acronyms

ACLU	American Civil Liberties Union
ADM	Algorithmic Decision Making
AGR	Automated Gender Recognition
AI	Artificial Intelligence
CAHAI	Council of Europe Ad Hoc Committee on Artificial Intelligence
ChatGPT	Generative Pre-Trained Transformer
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CRT	Civil Resolution Tribunal
EC	European Commission
ECHR	European Convention on Human Rights
EFF	Electronic Frontier Foundation
ESI	Electronically stored information
EU	European Union
FAIR	Findable, accessible, interoperable, and reusable
FRT	Facial recognition technology
FTC	Federal Trade Commission
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
HUDERAF	Human Rights, Democracy, and the Rule of Law Assurance Framework
vICCPR	International Covenant on Civil and Political Rights
ICESCR	International Covenant on Economic, Social, and Cultural Rights
IFIP	International Federation for Information Processing

ISO	International Organization for Standardization
IoT	Internet of Things
ITU	International Telecommunication Union
LAPD	Los Angeles Police Department
LLM	Large Language Model
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NDAS	National Data Analytics Solution
NGO	Non-governmental Organization
NIST	National Institutes of Standards and Technology
NLP	Natural Language Processing
OECD	Organization for Economic Co-operation and Development
SPC	Supreme People's Court
STF	Brazilian Supreme Court
SUPACE	Supreme Court Portal for Assistance in Courts Efficiency
TAR	Technology Assisted Review
UCL	Université Catholique de Leuven
UDHR	Universal Declaration of Human Rights
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
US	United States

WHY THIS TOOLKIT?

This Toolkit provides judicial operators with the knowledge and tools necessary to understand the benefits and risks of Artificial Intelligence ("AI") in their work. The Toolkit will support judicial operators in reducing potential human rights risks of AI by offering guidance on the relevant international human rights law instances, principles, regulations, and the emerging case law that underpin the use of AI responsibly.

The Toolkit responds to the UNESCO Recommendation on the Ethics of AI, adopted by 193 countries in 2021, that recommends "Member States should enhance the capacity of the Judiciary to make decisions related to AI systems as per the rule of law…".

What will you learn?

After studying the toolkit, judicial operators will be able to:

- Define AI and Algorithmic Decision Making (ADM) and understand them as socio-technical systems.
- Understand the key issues related to algorithmic bias and discrimination (such as gender bias, racial bias, and other intersecting forms of biases) and explain why these are important in judicial settings.
- Explain AI's impact on the following fundamental rights: privacy, freedom of expression, access to information, protection against discrimination, right to access to court, fair and impartial trials and hearings, and due process of law.
- Examine legal cases related to the use of AI, building upon their knowledge of the recent regulatory initiatives and case law related to algorithmic bias, inappropriate use of algorithms in decision-making.
- Apply tools like UNESCO's Ethical Impact Assessment for understanding the ethical impact of AI systems.

The Toolkit has four modules that complete a training programme on AI, human rights, and the rule of law for the Judiciary. The Toolkit also provides the necessary knowledge not only for judges but also for other actors involved in the dispute process, including lawyers and arbitrators.

- ***Module 1: Introduction to AI and the Rule of Law***

  Module one introduces the reader to the main concepts related to algorithmic governance, human rights, and the rule of law in the context of AI development. The Module defines terms such as AI, algorithms, algorithmic systems, and outlines their key characteristics and building blocks. Module one also discusses the importance of data and cybersecurity in the context of AI and provides an overview of the key risks associated with AI, such as black boxes.

- ***Module 2: AI Adoption in the Judiciary***

  Module two discusses AI adoption in the Judiciary. It outlines the uses of AI in the Judiciary, such as e-discovery and document review, use of generative AI to assist with the drafting of documents, predictive analytics and ADM support, risk assessment tools, dispute resolution, language recognition and analytics, digital file and case management. The Module then highlights case studies on AI deployment in the Judiciary in different countries and outlines the opportunities and challenges related to these use cases.

- ***Module 3: Legal and ethical challenges of AI***

  Module three presents key legal and ethical challenges related to AI in the Judiciary and summarizes the legal issues related to biometric identification and facial recognition technology. Module three discusses in detail the challenges related to AI and ethics based on the UNESCO 2021 Recommendation on the Ethics of Artificial Intelligence.[1]

- ***Module 4: Human rights and AI***

  Module four presents an in-depth analysis of human rights impacted by AI, such as (i) the right to access to court, fair trial, and due process, (ii) effective remedy, (iii) rights to protection against discrimination, (iv) freedom of expression and access to information, and (v) right to privacy and data protection The Module also gives an overview of key governance approaches to AI: risk based and human rights based.

Who will benefit from this Toolkit?

The Toolkit's primary target audience consists of judges, prosecutors, state attorneys, public lawyers, law universities and judicial training institutions.

How to use this Toolkit for teaching?

The Toolkit can be tailored to the specific needs of each judicial training programme. The number of hours and the duration of the training programme will depend on the methodology chosen by the judicial training programme. The programme may be taught as an online, classroom or hybrid learning scheme, and it may be offered as an intensive or a regular course of an undergraduate, a postgraduate, or a continuous education programme, based on the availability of trainers and/or the geographical distribution of the enrolled learners for a specific course, and the level of the accessibility and connectivity.

It is better to teach the programme as an organized effort to transfer the knowledge and develop the skills and attitudes that encourage actions geared towards promoting and protecting human rights in relation to AI.

1 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

Therefore, the following elements are recommended for any training based on the Toolkit:

- **Knowledge transfer:** In the context of this Toolkit, 'knowledge' refers to human rights standards and protection mechanisms that are pertinent to AI for the target group of learners. For example, in the context of a course where the target audience is judges, knowledge might refer to the human rights standards for deciding cases involving the use of AI.

- **Skills development:** A basic understanding of applicable human rights standards may be insufficient to enable learners to translate these norms into actual behaviour. The abilities are fine-tuned through practice, application and reflection, a process that can be initiated during the training through various activities but may need to be continued after the training course, including through adequately planned follow-up programmes. For example, the ability to conduct risk assessment of AI systems to determine whether they should even be deployed in the first place, rather than assuming deployment and then attempting ex post to mitigate harms.

- **Attitude development:** This involves acquisition and reinforcement of positive attitudes towards human rights and the rule of law, so that learners take action to promote and safeguard human rights in their everyday lives and professional responsibilities in adjudicating human rights violations involving ADM processes and AI.[2]

The training content has been made available online as an open resource and can be updated regularly by creating an online repository of presentations that trainers can refer to and re-use under creative commons open licenses (Attribution 4.0 International).[3]

2 Any changes leading to better respect for human rights - changes at the level of individual learners, their organization/group, and the larger community/society - that can plausibly be attributed to the training›s impact.

3 See: https://creativecommons.org/licenses/by/4.0/

Glossary

- **Aggregated data:** Data aggregation involves gathering a significant amount of information from a database and presenting it in a more manageable format.
- **AI as a "black box":** The term "black box" is used to denote a technological system that is inherently opaque, whose inner workings or underlying logic are not properly comprehended, or whose outputs and effects cannot be explained.
- **AI bias:** AI bias is a systematic difference in the treatment of certain objects, people, or groups (e.g. stereotyping, prejudice or favouritism) compared to others by AI algorithms.
- **Algorithm:** An algorithm refers to a series of instructions for performing calculations or other tasks, whether in mathematics or computer science. In the case of AI, an algorithm provides the instructions that enable a computer to learn how to learn from the environment and perform a set of tasks.
- **Algorithmic Decision Making (ADM):** Algorithmic decision making (ADM) refers to the use of 'outputs produced by algorithms to make decisions'.
- **Data Labelling:** Data labelling in Machine Learning (ML) is the process of recognizing raw data (pictures, text files, videos, etc.) and adding one or more relevant and useful labels to offer context for an ML model to learn from it. Labels may show whether a photograph contains a bird or an automobile, whether words were said in an audio recording, or whether an x-ray shows a tumour. Numerous application cases, including computer vision, natural language processing, and speech recognition, need data labelling.
- **Datafication:** The process of "datafication" refers to the proliferation of digital tools used to integrate, analyze, and display data patterns.
- **Data Trusts:** An independent organization that acts as a trustee for data providers and regulates the proper use of their data.
- **Deepfake:** A deepfake is any form of media (video, audio, or other) that has been altered or entirely or partially created from scratch.
- **Diffusion model:** Diffusion models are generative models that are more advanced than Generative Adversarial Networks (see below) on image synthesis. Most recently, Diffusion Models were used in DALL-E 2, OpenAI's image generation model and Google's Imagen.
- **Explainable AI (XAI):** Explainable AI (XAI) is defined as systems, algorithms, and models with the ability to explain their rationale for decisions, characterize the strengths and weaknesses of their decision-making process, and convey an understanding of how they will behave in the future.

- **Generative Adversarial Networks (GAN):** GANs are an unsupervised approach of deep learning that can generate hyper-realistic material. GANs are used for unsupervised deep learning techniques, such as generating realistic images or image datasets, performing text-to-image and image-to-text translations, ageing faces, and making emojis.
- **Generative AI:** Generative AI consists of Machine Learning (ML) algorithms that have been designed to create new content, including audio, code, images, text, simulations, and videos.
- **Hash Value:** Values returned by a hash function, which is used to convert digital data of arbitrary size into an output string with a fixed-size number of characters.
- **Human in the loop (HITL):** HITL refers to a process wherein an AI system is closely monitored by a human, who is responsible for making all final decisions. This is particularly important in fields like healthcare, where AI can provide invaluable support in making recommendations for cancer treatment, sepsis therapy, surgical planning, and more. While AI tools can help healthcare providers make informed decisions quickly and accurately, the ultimate responsibility for patient care always lies with the human expert.
- **Machine Learning (ML):** ML is a set of techniques that enables machines to learn automatically using patterns and deductions rather than direct instructions from a person. ML techniques frequently instruct machines to arrive at a result by providing numerous instances of correct results. However, they can also specify a set of guidelines and leave the machine to discover them on its own in the data.
- **Neural Networks:** Neural networks are a type of ML technique that enables computers to learn how to perform tasks by analysing training examples. Typically, these examples are pre-labelled. For example, an object recognition system may receive thousands of labelled images of objects such as cars, houses, and coffee cups. Through analysis, it can identify patterns in the images that correspond with specific labels. A neural network is designed to loosely resemble the structure of the human brain, with thousands or millions of interconnected processing nodes. These nodes are typically organized into layers, and data flows through them in a single direction, making them “feed-forward”. Each node receives data from nodes in the layer below it and sends data to nodes in the layer above it.
- **Natural Language Processing (NLP):** NLP is an ML technique that analyses vast amounts of human text or speech data (transcribed or acoustic) for specific properties, such as meaning, content, intention, attitude, and context.
- **Predictive analytics:** Predictive Analytics is the umbrella category of statistical tools and models, e.g., ML systems, that use and analyze historical data to create predictions about the future to guide decision making. These predictions can be low risk (e.g., which movie to recommend), medium risk (which loan application to propose accepting), or high risk (which defendant is most likely to engage in a particular behaviour).
- **Proxy discrimination:** Proxy discrimination in AI systems occurs when a seemingly neutral characteristic is substituted for a prohibited one.

- **Supervised Machine Learning:** Supervised machine learning involves providing a machine learning system with a set of data that is already labelled or classified, which the system can use to learn how to perform a particular task accurately according to the given instructions. The ML system is loaded with a dataset and the expected output. In the training phase, the ML model adjusts its variables to connect the inputs with the matching output. Creating a successful supervised learning algorithm requires a committed team of specialists to assess and scrutinize the outcomes. This involves data scientists who thoroughly examine the models produced by the algorithm to verify their precision against the source data and identify any inaccuracies caused by the AI.
- **Regulatory Sandboxes:** Regulatory tools allowing businesses to test and experiment with new and innovative products, services or businesses under supervision of a regulator for a limited period of time.

Technology
Innovation
SYSTEM
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Module 1

Introduction to AI and the Rule of Law

Module one introduces algorithmic governance, human rights, and the rule of law. It discusses definitions of AI, algorithms, and algorithmic systems, outlining their key characteristics and building blocks. The module underlines the importance of data and cybersecurity in the context of AI deployment in the Judiciary. It gives an overview of the key risks associated with AI deployment in the Judiciary, such as black boxes, and explains the human in the loop principle.

What will you learn?

After completing this module, the participants will be able to:

- Understand and explain key concepts related to AI, algorithmic governance, and the rule of law;
- Define and explain AI, algorithms, algorithmic systems, outlining their key characteristics and building blocks;
- Understand and recognize the risks associated with AI, such as black boxes and cybersecurity;
- Understand the importance of the human in the loop principle in the AI lifecycle;
- Understand why data is important in the context of AI.

1. Understanding AI and its building blocks

What are AI systems?

- As per UNESCO, AI systems are systems which have the capacity to process data and information in a way that resembles intelligent behaviour, typically including aspects of reasoning, learning, perception, prediction, planning or control.[4] In other words, AI systems are information-processing technologies that integrate models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments. AI systems are designed to operate with varying degrees of autonomy by means of knowledge modelling and representation, and by exploiting data and calculating correlations. AI systems may include several methods, such as (but not limited to):

- machine learning, including deep learning and reinforcement learning;

- machine reasoning, including planning, scheduling, knowledge representation and reasoning, search, and optimization.

It is important to note that such a definition would need to change over time, in accordance with technological developments. Further, AI is often used interchangeably with the term "machine learning" (ML), whereas AI is a much broader field that focuses on many things beyond ML, like knowledge representation, planning, and reasoning.[5]

In addition to the description above, Table 1 presents a snapshot of how different organizations define AI pragmatically, according to the set of tasks or functions the technology can undertake (OECD, ISO), or according to the humanistic ideals they seek to imbue into all manner of data-driven systems to ensure they contribute to the betterment of society (EC, ITU).

4 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

5 OECD (2019). Artificial Intelligence in Society, available at: https://www.oecd.org/en/publications/artificial-intelligence-in-society_eedfee77-en.html; Leslie D., Burr C., Aitken M., Cowls J., Katell M., and Briggs, M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, available at: https://ssrn.com/abstract=3817999 or http://dx.doi.org/10.2139/ssrn.3817999

Table 1. AI definitions in international and multilateral organizations

Organization	AI definition
OECD[6]	AI is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. When applied, AI has seven different use cases, also known as patterns, that can co-exist in parallel within the same AI system.
ISO[7]	Engineered system that generates outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives.
EC[8]	AI comprises systems that display intelligent behavior by analyzing their environment and taking actions—with some degree of autonomy—to achieve specific goals.
ITU[9]	AI refers to the ability of a computer or a computer-enabled robotic system to process information and produce outcomes in a manner similar to the thought process of humans in learning, decision-making, and problem-solving. In a way, the goal of AI systems is to develop systems capable of tackling complex problems in ways similar to human logic and reasoning.

AI systems in our daily life

AI is already part of our daily lives, whether we know it or not. Think of your email inbox: You may notice that certain emails end up in your spam folder, while others are categorized as "social" or "promotion." How does this happen? Did you know that Google has implemented AI algorithms to automatically categorize and filter emails? These algorithms are programs trained to identify specific elements within an email that indicate it might be spam. When the algorithm recognizes these elements, it marks the email as spam and moves it to your spam folder. While the algorithms aren't perfect, they are constantly being improved. If you happen to find a legitimate email in your spam folder, you can let Google know that it was wrongly labelled as spam. This feedback helps to enhance the accuracy of the algorithm.[10]

Another example of an AI system in our daily interactions is in the form of a customer service chatbot. When you type in your question, the chatbot uses an algorithm to recognize keywords and determine the type of assistance you require. Based on the existing and newly acquired information, the machine learning model generates an appropriate response. As the chatbot interacts with more customers and receives additional data, it improves over time.[11]

6 OECD (2019). Artificial intelligence and responsible business conduct, available at: https://mneguidelines.oecd.org/RBC-and-artificial-intelligence.pdf
7 ISO (2021). ISO/IEC DIS 22989, available at: www.iso.org/standard/74296.html
8 European Commission (2018). Communication Artificial Intelligence for Europe, available at: https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe
9 ITU (2018). Policy Considerations for AI Governance, available at: https://www.itu.int/en/ITU-T/studygroups/2017-2020/03/Documents/Shailendra%20Hajela_Presentation.pdf
10 See: https://dig.watch/technologies/artificial-intelligence
11 Bravo K. (2023). How Does AI actually work?, available at: https://blog.mozilla.org/en/internet-culture/how-does-ai-work/

Other examples of everyday AI systems include Netflix's recommendation engine for suggesting films and TV shows based on our preferences, or voice assistants like Siri and Alexa that help us with simple queries.

Activity: Questions for Reflection

1. What comes into your mind when you hear the term AI? List your connotations freely and compare them with a peer. Did you come up with any similar ideas? How are these ideas possibly reflected in dominant public discourses on AI?

2. Envision the technological development of the future three decades in the following environments (alternatively, pick only one of them): home/family, school, healthcare. Which processes have been automated? How has automation affected people's behaviour, social interaction and experiences?

Invite the training participants to watch the following videos.

Source: BBC, https://youtu.be/fvtrRGmv7aU

Source: OECD, https://youtu.be/6Y_ysDHn4uU

What is an algorithm?

An algorithm refers to a series of instructions for performing calculations or other tasks, whether in mathematics or computer science. In the case of AI, an algorithm provides the instructions that enable a computer to learn how to learn from the environment and perform a set of tasks.[12]

While a general algorithm can be simple, AI algorithms are more complex.

AI algorithms are designed to learn from training data, which can either be labelled or unlabelled. The algorithm uses this information to enhance its capabilities and carry out its tasks. Some AI algorithms are capable of continuous learning and can incorporate new data input to refine their process, while others require the intervention of a programmer to optimize their performance.[13]

Algorithmic decision making (ADM) refers to the use of "outputs produced by algorithms to make decisions."[14]

> Algorithms work by taking a set of inputs, such as a person's age, district of residence, marital status, or income, and running them through a set of steps that create an output or outputs, or decisions, for that person or group such as eligibility for a financial assistance program or the public school to which a child is assigned. Algorithms are used across various sectors and purposes, from healthcare decisions, public benefits eligibility, infrastructure planning, budget allocation, among other sectors, with varying degrees of complexity and inputs.

Waves of AI development

The first wave AI systems were expert or rules-based systems, where a computer followed specific programming to generate outputs. However, the second wave AI systems, based on machine learning, learn from the training data and infer rules to predict specific outcomes. The third-wave AI systems combine the advantages of the previous two waves and have added capabilities of being able to respond to the context in which they are used and provide users with explanations for their decision-making process.[15]

The sections below explain and focus on (i) expert systems and traditional programming and (ii) machine learning.

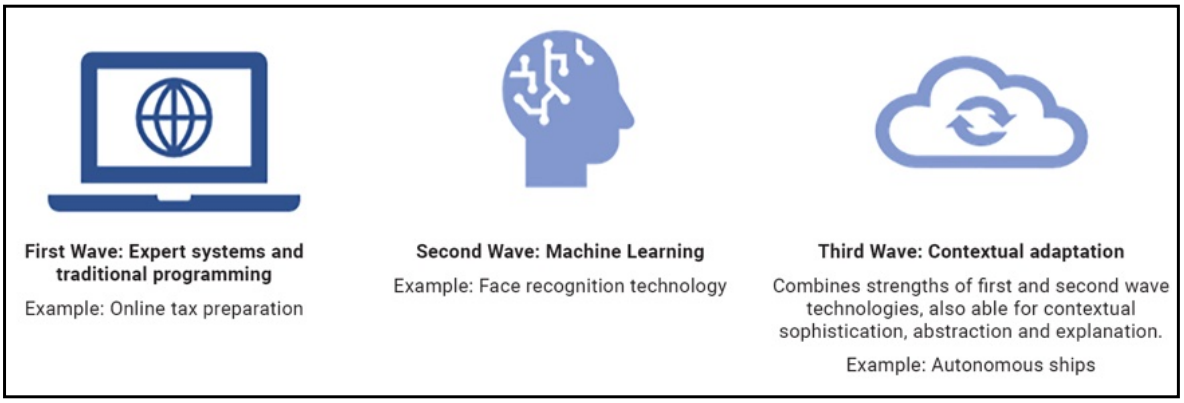
12 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4162985 Statement on Algorithmic Transparency Accountability, Association for Computing Machinery (2017), available at: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf; also see: https://www.tableau.com/data-insights/ai/algorithms.

13 OECD (2019). Artificial Intelligence in Society, available at: https://www.oecd.org/en/publications/artificial-intelligence-in-society_eedfee77-en.html

14 Access Now (2018). Human rights in the age of artificial intelligence, available at: https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdf

15 GAO (2021). Artificial Intelligence, An Accountability Framework for Federal Agencies and Other Entities, available at: https://www.gao.gov/products/gao-21-519sp.

Figure 1. Waves of AI development



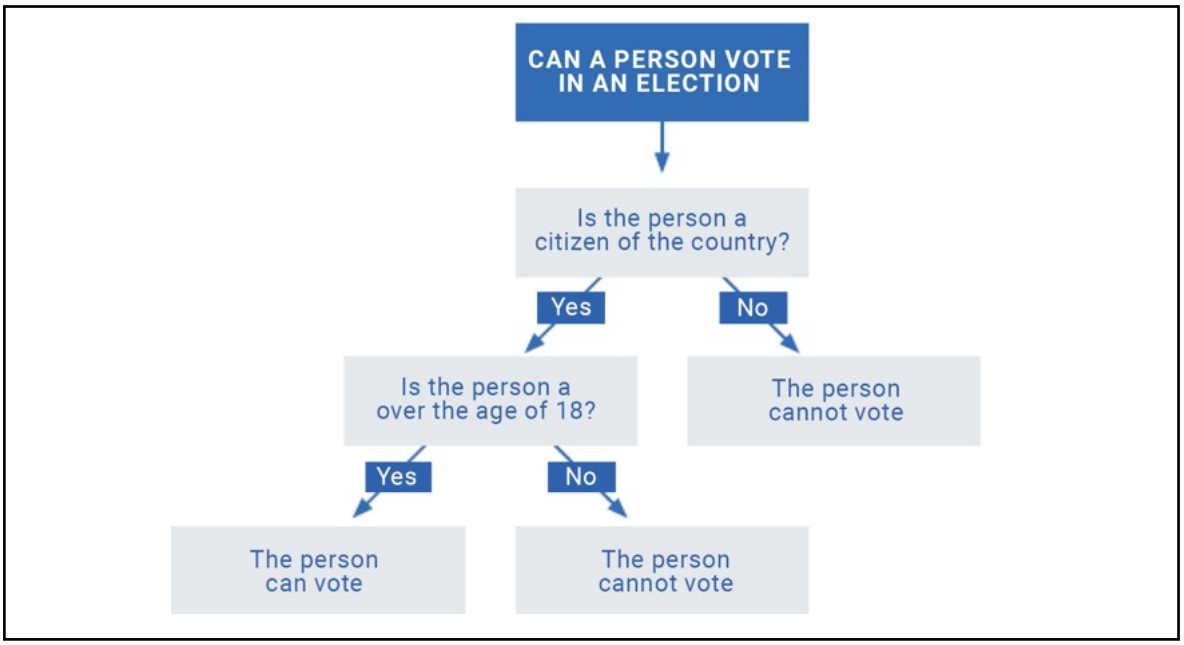
Source: Adapted from GAO (2021). Artificial Intelligence, An Accountability Framework for Federal Agencies and Other Entities, available at: https://www.gao.gov/products/gao-21-519sp

Expert systems and traditional programming

An "expert system" is a "first-generation" AI system that makes forecasts, recommendations, or conclusions based on data input. It involves a sequence of clearly programmed stages and so-called "if...then" rules, which a computer can apply to produce an output. These systems are typically incapable of dealing with fresh information or unexpected challenges.

The possible choices are referred to as "nodes" in a decision tree, which is a visual representation of the rules of the expert system. Figure 2 below shows an example of a decision tree that decides whether a person may vote in an election in a country where the only prerequisites for being able to vote are that the individual be over 18 and a citizen of a particular country. Given that each branch only has two nodes, Figure 2 is an example of a "binary" decision tree.

Figure 2. Example of a decision tree



Source: Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, available at: https://ssrn.com/abstract=4162985

First generation AI expert systems are widely used in planning and optimization systems. Among others, examples include tax processing software, customer service and technical support systems, and medical diagnosis systems. Another example is a fraud alert method where an expert specifies that if the supplied administrative information has more than five inaccuracies, the system should issue an alert indicating that this dossier should be investigated.

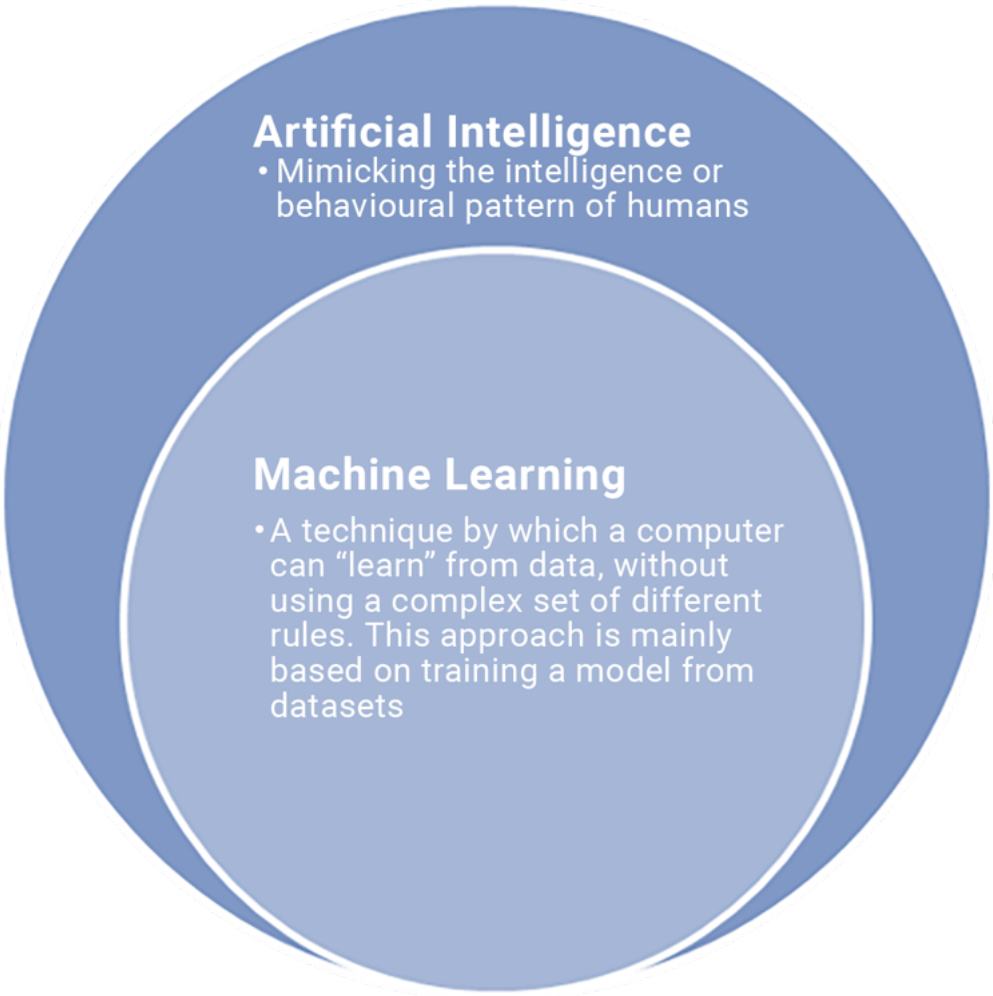
Initially, mastering a programming language was necessary to create rules in a language that a machine could understand. The concept behind an "expert system" was that the rules might be developed by a subject-matter expert (e.g., lawyer) who did not possess programming abilities. A variety of "no-code" platforms are now available that make it simple to "program" a computer to follow a certain procedure or come to conclusions based on a set of rules. Examples of such platforms include Austlii's Datalex[16], Josef Legal[17], Checkbox[18], Neota Logic[19] and Realta Logic[20]. These platforms allow legal professionals to design a set of rules using, depending on the platform being utilized, words, statements, arrows, drag-and-drop or drop-down menus, or other similar processes, depending on the platform being utilized. As a result, even a lawyer without programming experience can encode a decision tree like the one shown in Figure 2.[21]

What is Machine Learning?

AI systems increasingly employ machine learning (ML), which is a subset of AI. ML is a set of techniques that enables machines to learn automatically using patterns and deductions rather than direct instructions from a person.[22] ML techniques frequently instruct machines to arrive at a result by providing numerous instances of correct results. However, they can also specify a set of guidelines and leave the machine to discover them on its own.[23]

16 See: https://austlii.community/wiki/DataLex
17 See: https://joseflegal.com/
18 See: https://www.checkbox.ai/
19 See: https://neota.com/
20 See: https://www.realtalogic.com/
21 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4162985
22 OECD (2019). Artificial Intelligence in Society, available at: https://www.oecd.org/en/publications/artificial-intelligence-in-society_eedfee77-en.html
23 Ibid. Numerous methods that have been employed by economists, scientists, and engineers for years can be found in ML. These include principal component analysis, decision trees, deep neural networks, and linear and logistic regressions. See: https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57/cb6d9eca-en.pdf

Figure 3. The relationship between AI and ML



Source: Authors

There are many ML applications. Some are designed for a specific problem, like speech or image recognition, while others can be used for a wider range of tasks.[24] ML has been integrated into products to tackle a variety of issues that are too complicated for "first-generation" AI systems or human decision-making. ML powers chatbots, predictive text, language translation apps, Netflix recommendations, and the organization of social media feeds. It also enables self-driving cars and machines capable of diagnosing medical conditions using image analysis.[25]

ML systems "learn" as they analyse data. ML is distinct from human learning. While seeing few photographs of a cat will enable an average child to comprehend the term "cat" and recognize additional images as cats, ML systems require a far larger data set to perform the same categorization task. The ML program relies on a database containing cat and dog images.

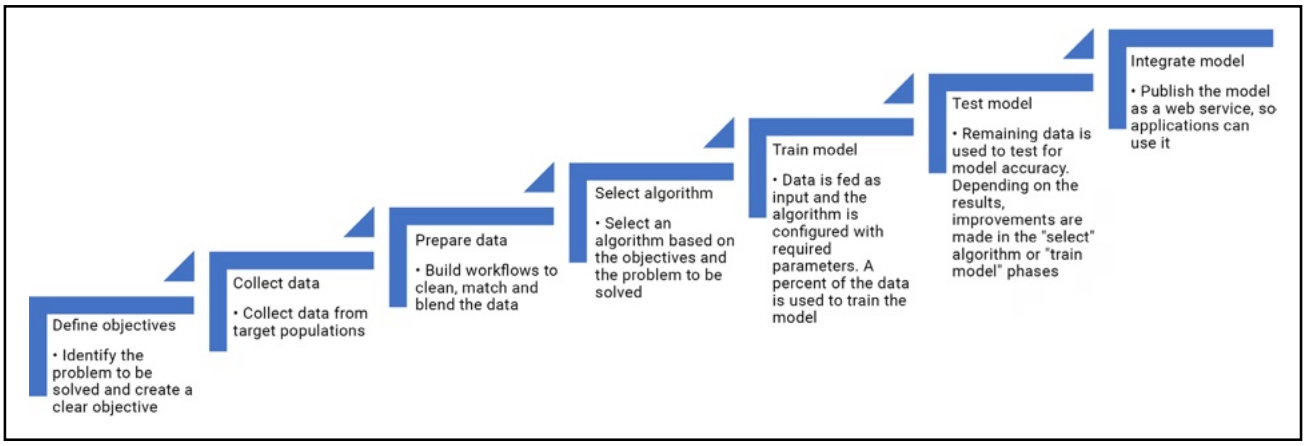
24 OECD (2019). Artificial Intelligence in Society, available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/02/oecd-framework-for

25 Brown S. (2021). Machine learning, explained, available at: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Each image is labelled with "cat" or "dog". If the ML program is shown enough labelled pictures, the ML program will start to differentiate the characteristics of each animal (ML training or fitting). Once the ML program learns, it will be able to guess which class each picture belongs to. Very similar experiments can be conducted with text.[26] Another good example of an ML program is the process of assigning credit scores by financial institutions, where the data used to train the ML system is already classified as positive or negative depending on the customer's credit history.[27] We need to remember that the efficacy of ML models depends on the quantity of training data available, the quality of training and input data, and the amount of computing power used to build the model.[28]

Figure 4 below gives a simplified overview of an ML process, consisting of the following phases: (i) definition of objectives; (ii) data collection; (iii) data preparation; (iv) selection of the algorithm; (v) training of the model; (vi) testing of the model; and (vii) model integration.

Figure 4. Simplistic overview of the ML process



Source: Authors

26 Medvedeva M., Vols M., Wieling M. (2020). Using machine learning to predict decisions of the European Court of Human Rights, Artif Intell Law, 28, 237–266, available at: https://link.springer.com/article/10.1007/s10506-019-09255-y

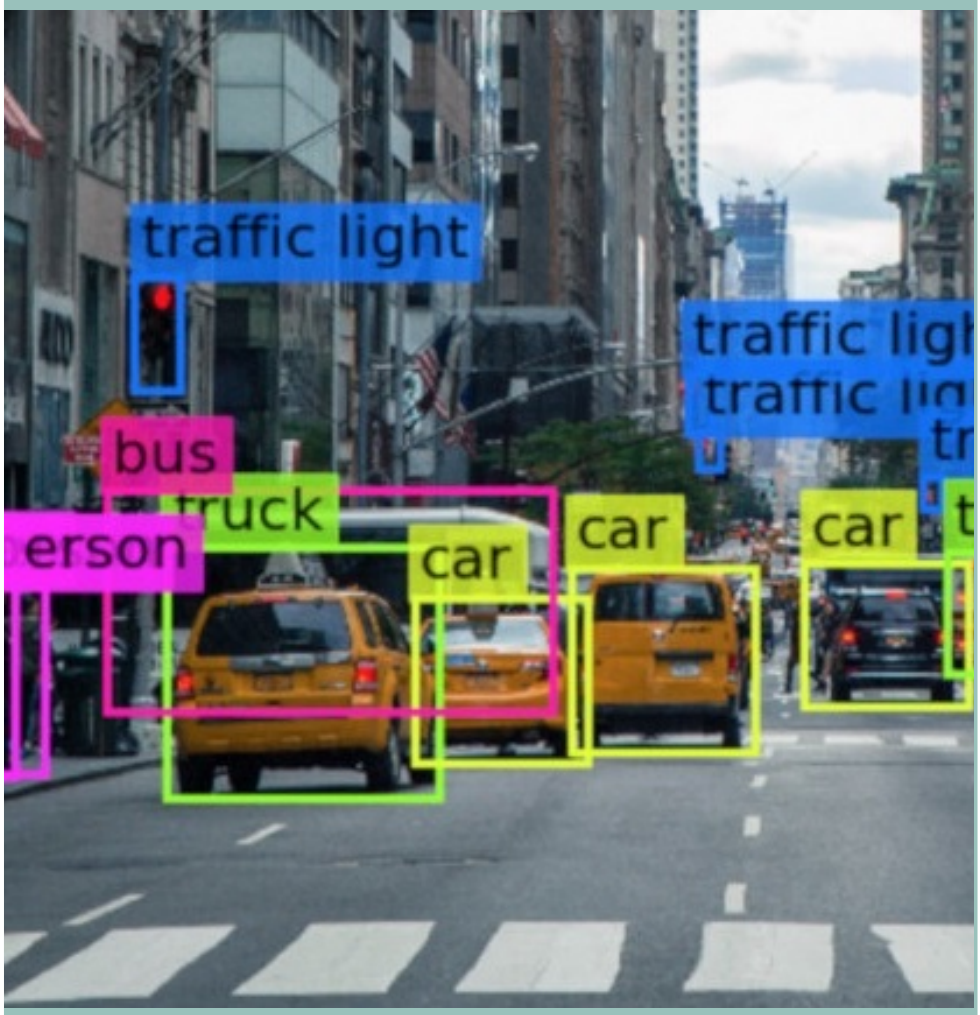
27 The Royal Society (2012). Machine Learning: The Power and Promise of Computers that Learn by Example, available at: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf 16; Allens Linklaters (2018). AI Toolkit: Ethical, Safe, lawful; Practical Guidance for AI Projects, available at: https://www.linklaters.com/fr-fr/knowledge/publications/alerts-newsletters-and-guides/2018/november/global---ethical-safe-lawful-a-toolkit-for-artificial-intelligence-projects 9; https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-final-report-2021.

28 Stankovich M., Behrens E., Burchell J. (2023). Toward Meaningful Transparency and Accountability of AI Algorithms in Public Service Delivery, disponible en: https://www.dai.com/uploads/ai-in-public-service.pdf

Data labelling in ML

Data labelling in ML is the process of recognizing raw data (pictures, text files, videos, etc.) and adding one or more relevant and useful labels to offer context for an ML model to learn from it. Labels may show, for instance, if a photograph contains a bird or an automobile, whether words were said on an audio recording, or whether an x-ray shows a tumour. Numerous application cases, including computer vision, natural language processing, and speech recognition, need data labelling[29].

Figure 5. Data labelling in ML



Example of data labelling.

Source: Energy (2021). The One, Two, Threes of Data Labeling for Computer Vision, available at: https://medium.com/unpackai/the-one-two-threes-of-data-labeling-for-computer-vision-4c0b022cef4

29 See: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/

The discovery process in litigation can serve as a great example of showcasing the complexity of using ML in the Judiciary. See Figure 6 below.

Figure 6. Discovery in litigation: Three possible levels of automation

1 level - No automation at all. At this stage, a paralegal sifts through the legal documents following a predetermined set of parameters.

II level - Automation without ML. A computer system uses fixed criteria, such as date range, lists of phrases, location of files, to conduct the discovery of documents.

III level – ML. A paralegal labels which documents will be part of the discovery (this is the training data). Then, an ML system may be used to infer the searchability criteria based on patterns in the human- labelled training data rather than using only a human to identify which characteristics are necessary for discoverability. The trained ML model will classify the remaining documents into those that are and are not likely to be discoverable using these patterns.

Source: Authors

Activity: Training participants discuss the following hypothetical scenario about the use of AI-generated evidence in judicial proceedings. What would you do if you were in a similar situation? What key legal issues will you take into consideration?

In the not-so-distant future, AI-generated evidence plays a pivotal role in a high-profile court case. Here's how it unfolds:

Case Background: A prominent tech company is accused of using biased algorithms in their hiring process, resulting in discrimination against certain demographic groups. The case has garnered significant public attention and is being closely watched for its potential implications on AI ethics and corporate responsibility.

AI-Generated Evidence:

1. **Algorithmic Audit Report:** The plaintiffs have employed a team of AI ethicists and data scientists to conduct a comprehensive audit of the company's hiring algorithms. They present a detailed report generated by AI systems that highlights instances of bias and discrimination in the algorithm's decision-making process.
2. **AI-Generated Simulation:** To demonstrate the algorithm's behavior, the plaintiffs introduce an AI-generated simulation that mimics the company's hiring process. This simulation uses historical data to show how the algorithm tends to favor certain demographic groups over others.
3. **AI-Generated Expert Testimony:** The defense calls an AI ethics expert who uses natural language processing AI to analyze the company's internal communications and documents. The AI identifies instances where employees expressed concerns about algorithmic bias, potentially suggesting that the company was aware of the issue.

Legal Implications: The introduction of AI-generated evidence presents several legal challenges and considerations:

1. **Admissibility:** The court must determine the admissibility of AI-generated evidence, assessing its reliability and relevance to the case.
2. **Expert Testimony:** The court grapples with the question of whether AI can be considered an "expert witness" and how its testimony should be treated.
3. **Ethical Implications:** The case raises ethical questions about the responsibility of companies when deploying AI systems and the potential consequences of algorithmic bias.
4. **Impact on Precedent:** The outcome of this case could set a precedent for how AI-generated evidence is treated in future legal proceedings, influencing the legal landscape regarding AI ethics.
5. **Human Oversight:** Despite AI-generated evidence, human judgment remains crucial in interpreting the evidence, ensuring fairness, and making legal decisions.

This hypothetical scenario underscores the evolving role of AI in legal proceedings, as well as the need for robust legal frameworks to address the complexities and ethical concerns associated with AI-generated evidence in courtrooms.

2. Why is data important in the context of AI?

AI algorithms require access to data—machines cannot "learn" unless they have large datasets from which to discern patterns. The availability of data is a necessary requirement for the development of AI allowing it to do certain tasks previously performed manually by humans.

The process of "datafication" refers to the proliferation of digital tools used for the integration, analysis, and display of data patterns. Datafication indicates that numerous aspects of social life assume the form of digital footprints. Friendships become "likes" on Facebook, movements across the city leave vast digital imprints in GPS-enabled gadgets, and information searches reveal what individuals and communities value or desire.[30]

Once Internet-connected devices start communicating with one another, an extraordinary quantity of new data is delivered unknowingly and virtually unnoticed by most users. For example, there are metadata (data about data), such as the routing information contained inside the headers of emails or text messages, or the geolocation information concealed within a digital photograph. Metadata, as structured information, can be more easily compared and evaluated by algorithms, and can, therefore, frequently give unusually exact information on the interests, movements, and relationships of individuals.

Digital platforms have access to a lot of information about what people are doing online. These massive streams of digital traces, called big data, can be used in conjunction with automated sorting techniques, such as algorithms and AI, to reveal important patterns and lead to analytical insights on customers, diseases, and criminal activities. Many digital platforms and firms seek to lock in customers early on by becoming the place where people buy books or stream movies, for instance. They also want to build closed-off ecosystems, like Netflix or Amazon, where they can control and extract value from data.[31]

The quality of data impacts the outcome of AI in terms of bias [for AI bias, refer to Module 3]. Data should ideally be free of bias, data ownership must be clearly established, and algorithms must be transparent enough to indicate stakeholders' liability. The obligations of all stakeholders in the AI lifecycle must be defined to prevent damage and repair or compensate for harm caused by AI systems.

When deciding cases that involve AI deployment and its impact on human rights, judicial operators should consider the following questions related to data and datasets that feed into AI systems (see Table 2).

30 Matteson A. (2018). The Concept of Datafication; Definition & Examples, available at: https://www.datasciencecentral.com/the-concept-of-datafication-definition-amp-examples/

31 Flyverbom M., Deibert R., Matten, D. (2019). The Governance of Digital Technology, Big Data, and the Internet: New Roles and Responsibilities for Business. Business & Society, 58(1), 3–19, available at: https://doi.org/10.1177/0007650317727540

Table 2. Questions related to data and datasets that feed into AI systems.

Questions	Issues to consider
Data access and availability	The absence of necessary systems that generate and maintain robust, accurate, and relevant data has made the development of AI applications challenging in some contexts.
Data accuracy	Access to accurate data is crucial for successfully deploying AI and digital resources. A good practice in safeguarding data accuracy is the practice of the called «algorithmic disgorgement» that requires AI system developers to remove any data that was obtained illegally and used to train the AI systems.[32]
Data quality	One of the key impediments to the effective deployment of AI in the Judiciary is access to FAIR (findable, accessible, interoperable, and reusable) data. This problem is exacerbated in certain contexts because data are not always digitized and not easily accessible. Key questions to ask in this regard: What is the quality of data that the AI system is trained on? Is there a risk of data bias and amplification of incorrect information using AI? The problem is that data that feeds AI systems can be inaccurate, incomplete, or contain errors or unimportant material. Data might be infused with bias. Many times, machines are already collecting skewed data that comes from erratic and biased reality. For instance, clinical trials often exclude women and people of colour, leading to inadequate data representation. This could have severe consequences if algorithms trained using such data are used to analyze skin images or prioritize care for patients. As a result, it is crucial to ensure that AI algorithms are trained using representative data to avoid such biases and ensure equitable outcomes for all.[33] Example: most AI systems used in criminal justice are statistical models, based on law enforcement or criminal records data that represent structural biases and social inequalities. This data is a record of the crimes, locations, and policed groups, and is not a necessary record of the actual occurrence of crime. This data used in AI systems can reinforce and re-enter patterns of discrimination in justice or law enforcement systems.[34] Regulators of AI models should ensure that the data used adheres to the FAIR principles and is collected ethically before certifying the model as fit for the market. This could be further supplemented by organizational quality assessment in pre-market checkpoints. These conditions can signal to the industry that data integrity and ethical collection are of paramount importance placing AI solutions on the market and lead to positive structural changes in how enterprises function.

32 The FTC used this practice to compel Everalbum, creators of the now-defunct app Ever, to delete facial recognition systems that were developed using content obtained from the app›s users. See also: Kay K. (2021). Why the FTC is forcing tech firms to kill their algorithms along with ill-gotten data, available at: https://digiday.com/media/why-the-ftc-is-forcing-tech-firms-to-kill-their-algorithms-along-with-ill-gotten-data/

33 Siwicki B. (2021). How does bias affect healthcare AI, and what can be done about it?, available at: https://www.healthcareitnews.com/news/how-does-bias-affect-healthcare-ai-and-what-can-be-done-about-it

34 Fair Trials (2021). Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU Policy Paper, available at: https://www.fairtrials.org/app/uploads/2022/01/Regulating-Artificial-Intelligence-for-Use-in-Criminal-Justice-Systems-Fair-Trials.pdf

Questions	Issues to consider
Data representativeness	A dataset is representative if it accurately reflects or measures the population or phenomenon it is meant to record, relative to its intended application.[35] Example: Excessive reliance on "automated" data collection techniques can leave out extremely vulnerable groups and erode trust in automated decision making. People without digital access (i.e., those without connectivity or devices) or who lack digital skills will not be considered in analyses of the population and their requirements. Digital divides in many Global South countries have led to "data invisibility," which is likely to impact historically marginalized groups like women, castes, tribal communities, religious and linguistic minorities, and migrant labour. The usefulness and validity of AI algorithms developed on readily available data may be constrained by biases perpetuated by data invisibility. This underlines the requirements for algorithmic transparency and accountability.
Data ownership	A key issue in AI development and deployment is data ownership, i.e., who owns, manages, and collect the data that goes into the AI system. Important issues to consider in this regard are defining the goals (why do we need the AI system) and determining what training data to acquire and how to categorize the data. Therefore, human judgements are constantly required when compiling datasets and developing algorithms for prediction.
Data storage/data minimization	Long-term storage of personal data entails hazards, as data are susceptible to exploitation in ways that were not anticipated at the time of data collection. The data may become outdated, irrelevant, or contain historical misinterpretation over time, which could lead to skewed or incorrect results from data processing in the future.[36]
Data protection/ privacy	Adequate data protection laws tackle issues such as data privacy (a basic human right), data management and sharing and innovative mechanisms for data governance such as data sandboxes and data trusts. Current data policies and regulations among countries and regions are highly fragmented, with diverging global, regional, and national regulatory approaches. Many countries and regions have taken steps to update rules on the use of personal data. The EU General Data Protection Regulation[37] (GDPR) imposes a long list of requirements for companies processing personal data. Violations result in fines that could total as much as 4 % of global annual turnover. The GDPR enables better control over personal data, entitling individual protection of anonymity, pseudonymity, and the right to be forgotten. Data portability gives individuals the right to request that their data be transferred to another controller and for data controllers to use common formats. More than 30% of countries, primarily developing ones, have no data governance legislation, and few have developed comprehensive data protection law[38]. Other regional frameworks for setting rules on privacy of personal data include the APEC Privacy Framework (2015); and the OECD's Privacy Guidelines (2013) and the Council of Europe's Convention 108+[39], which has updated guidelines on data protection. It is worth nothing that in countries that do not have a data protection system in place, the courts may have to lay down guidelines for the use of data, which would be in consonance with legal rights.

35 AAAS. Artificial Intelligence and the Courts: Materials for Judges, available at: https://www.aaas.org/ai2/projects/law/judicialpapers
36 UN Human Rights Council (2021). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, availableat:https://www.ohchr.org/en/documents/thematic-reports/ahrc4831-right-privacy-digital-age-report-united-nations-high
37 Complete guide to GDPR compliance, available at: https://gdpr.eu/
38 UNCTAD, Data Protection and Privacy Legislation Worldwide, available at: https://unctad.org/page/data-protection-and-privacy-legislation-worldwide
39 Council of Europe, Modernisation of Convention 108, available at https://www.coe.int/fr/web/data-protection/convention108/modernised

Questions	Issues to consider
Data infrastructure	Today's progress in AI and big data is fueled by better digital connections, growing amounts of data, sophisticated algorithms, and increased processing power. AI and big data can greatly improve lives in developing countries and help achieve the UN Sustainable Development Goals. Policymakers should aim to enable, incentivize and/or accelerate investment in building adequate and affordable data infrastructure. Investment in software, hardware, and broadband connectivity is needed for widespread data access and use. This is critical for reaching the underserved. Incentivizing the creation of FAIR data and FAIR data infrastructure is crucial.[40]
Additional questions to ask	• Did the AI system undergo algorithmic transparency audits or privacy impact assessments? • Were there any privacy enhancing techniques used to preserve privacy of the data? • What is the status of information and cybersecurity for data privacy?

3. AI systems as "black boxes"

The term "black box" is used to denote a technological system that is inherently opaque, whose inner workings or underlying logic are not properly comprehended, or whose outputs and effects cannot be explained.[41] Many AI systems are considered to be "black boxes,", i.e., highly complex systems whose decision-making and reasoning processes are not easily understood by users, and even by their developers. This can make it extremely difficult to detect flawed outputs, particularly in AI systems that discover patterns in the underlying data in an unsupervised manner.

AI systems analyse the training data to identifying complex patterns and then learn these patterns to classify new data that they may be fed. Many AI systems do not, however, explain how data could be interrelated and how they reach a certain decision or predict a certain outcome. These systems can be far too complex for human comprehension, even for those who program and train them.[42] They evolve and learn continuously and have unpredictable behaviour. They may be able to deduce facts and correlations from proxy variables, such as purchase history or geography.

40 FAIR Principles, available at: https://www.go-fair.org/fair-principles/
41 AAAS, Artificial Intelligence and the Courts: Materials for Judges, available at: https://www.aaas.org/ai2/projects/law/judicialpapers
42 OECD, AI in Society, available at: https://www.oecd-ilibrary.org/sites/969ff07f-en/index.html?itemId=/content/component/969ff07f-en

In depth: Proxy discrimination in AI systems

Proxy discrimination in AI systems takes place when a seemingly neutral characteristic is substituted for a prohibited one.[43]

For example, financial institutions often use postal codes and neighbourhood limits (geography), this data may capture the race of loan applicants, since some postal codes may be associated with low-income social groups, ethnic or racial minorities. Similarly, an AI system created by an insurance company may increase premiums for applicants who may be members of a Facebook group dedicated to improving the availability of cancer predicting genetic testing. Under these conditions, the insurer is probably engaging in indirect genetic discrimination by using proxies, such as demand for a certain kind of genetic testing and membership to a specific Facebook group, to deduce the link between these proxies and genetic history (a controversial practice) and charge higher insurance premiums to such individuals.[44] Another example would be proxies related to age, "twenty years of professional experience" indicates that the person must be at least in their mid-forties.

The rights to privacy and non-discrimination in automated decision-making systems call for data minimization, limitation, or prohibition of certain uses of data, or data removal (refer to Table 2 above). However, an AI system may make a prediction based on proxy data that has a close resemblance to the restricted categories of data. In addition, the only way to discover these proxies is to acquire sensitive or private information such as race. If such data are acquired, it becomes crucial to guarantee that they are used exclusively for adequate and legitimate purposes.[45] For instance, even though algorithm creators may have made a conscious effort to prevent racial bias by excluding race as a parameter, the algorithm will nevertheless produce results that are racially biased if it includes typical proxies for race, such as income, education, or postal code.

The opacity of AI algorithms and the difficulty in determining liability for the decisions produced by AI systems mean that human rights harms can occur, and no responsibility is fixed for these harms. Without incorporating ethical and human rights safeguards in AI design and deployment, the risks related to AI will intensify. This will have an impact on deepening existing inequalities embedded in datasets used to train algorithms. For instance, these inequalities might stem from such the bias of the developers. This will severely and disproportionally affect underprivileged, underserved, and marginalized groups, and those who are subject to intersecting forms of discriminations.

43 Downs J., Auchterlonie S. (2022). Proxy Problems—Solving for Discrimination in Algorithms, available at: https://www.bhfs.com/insights/alerts-articles/2022/proxy-problems-solving-for-discrimination-in-algorithms

44 Iowa Law Review (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data, available at: https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data

45 O'Neil C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York: Crown.

Another problem is the misuse of intellectual property safeguards. Algorithmic tools often fall under the shield of proprietary software and trade secrets claims to protect the technology behind the algorithms from outside scrutiny (see People v. Chubbs discussed in Module 4 below). This practice might impede any defence effort to challenge the reliability of the science underlying the AI tool. When AI systems are used in operations on behalf of stakeholders of the justice system, there is an accentuated need for accountability, transparency and explainability. Intellectual property safeguards of the data and the algorithmic system may prevent such transparency and accountability. AI governance stakeholders will need to find a balance between transparency as part of AI ethics and the legitimate need to protect commercial secrets when private companies develop AI tools.

Activity: Trade secrets, algorithms and fundamental rights: the case study of the Educational Value-Added Assessment System (EVAAS) algorithm

Trade secrets that protect algorithms affect fundamental rights. Read the case study below and discuss how a similar case would be judged in your country. How would this case be decided under your national laws?

Between 2011 and 2015, Houston teachers' work performance was evaluated using a "data-driven" algorithm – EVAAS. The programme enabled the board of education to automate choices over whether teachers were granted bonuses, penalized for poor performance, or even terminated. The source codes are trade secrets owned by SAS, a third-party vendor. As such, the teachers were unable to contest the decisions or receive an explanation for how the EVAAS reached its decisions.

A lengthy civil litigation occurred, and in 2017, a US federal judge concluded that the instructors' constitutional rights were violated by the deployment of the secret algorithm to evaluate employee performance without appropriate explanation. The judge had to strike a balance between the understandable right of the private vendor to preserve its trade secrets and the teachers' constitutional right to due process, which protects US citizens from deprivations of life, liberty, or property that are fundamentally unjust or erroneous.

The court decision stated that the teachers and the Houston Federation of Teachers must be able to independently check and contest the evaluation results produced by the algorithm. However, SAS declined to reveal how their EVAAS algorithm operates internally. As a result, the Houston school system no longer uses the EVAAS algorithm.

Source: Hung K-H., Liddicoat J. (2018). The future of workers' rights in the AI age, available at: https://policyoptions.irpp.org/magazines/december-2018/future-workers-rights-ai-age/

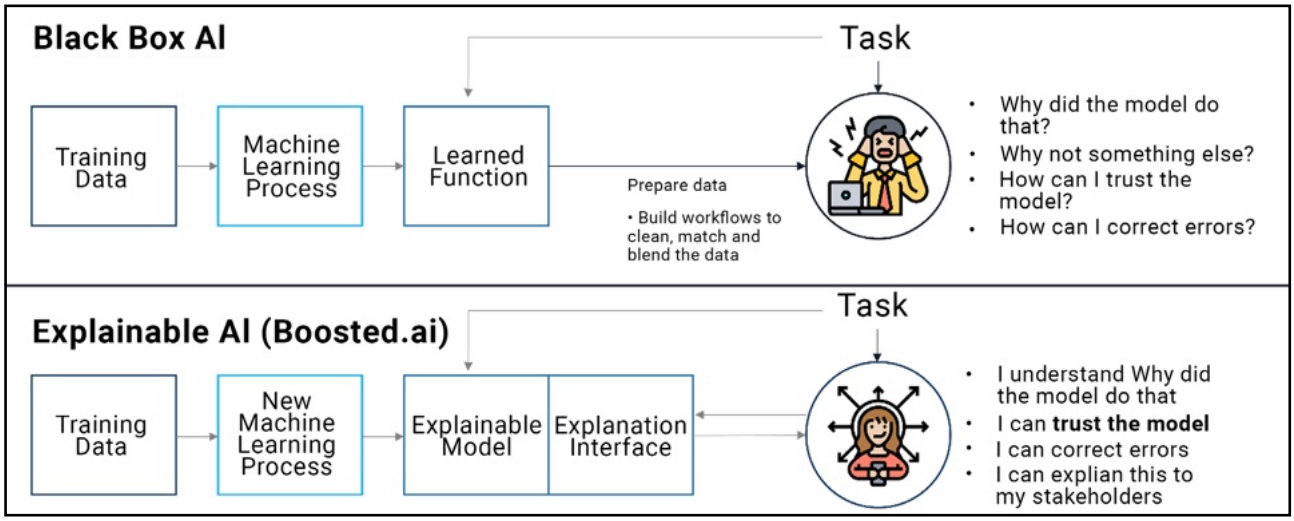
Explainable AI (XAI)

The discussion around black box aspects of AI systems is continuously evolving. The advancements in AI research have led to the development of AI models that are not black boxes.

Explainable AI is defined as systems, algorithms, and models with the ability to explain their rationale for decisions, characterize the strengths and weaknesses of their decision-making process, and convey an understanding of how they will behave in the future.

Researchers in XAI concentrate on creating AI models that can be comprehended by people, as well as producing explanations of ML outputs that are usable. This audience should have the opportunity to analyze the generated model and discern its meaning, i.e., to understand the structure of the system.

Figure 7. Black box AI versus explainable AI



Source: https://boosted.ai/

For example, Angelino et al (2018) developed an interpretable ML model for forecasting re-arrest that only includes a few rules on an individual's age and criminal history. The complete ML model predicts that a person will be rearrested within two years after their evaluation if they have committed three or more past offences, are between the ages of 18 and 20 and male or are between the ages of 21 and 23 and have committed two or three prior offences. This set of guidelines is as accurate as the widely used (and proprietary) COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) black box model, which is used in Broward County, Florida. Please refer to the section on algorithmic bias to get acquainted with COMPAS.

Case Study: The US National Institutes of Standards and Technology (NIST) guidance on AI explainability

The US NIST has issued guidance on AI explainability that might be part of impact assessment systems. The NIST draft guidelines suggest four principles for explainability for audience sensitive, purpose driven, automated decision-making systems (ADSs) assessment tools: (1) Systems offer accompanying evidence or reason(s) for all outputs; (2) Systems provide explanations that are understandable to individual users; (3) The explanation correctly reflects the system's process for generating the output; and (4) The system only operates under conditions for which it was designed or when the system reaches sufficient confidence in its output. These four principles shape the types of explanations needed to ensure confidence in algorithmic decision-making systems, such as explanations for user benefit, for social acceptance, for regulatory and compliance purposes, for system development, and for owner benefit.

Source: NIST (2020). Four Principles of Explainable Artificial Intelligence, available at: https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence

4. The human in the loop principle

Realizing that many AI systems are black boxes and prone to bias, judicial operators will start addressing questions concerning the extent to which humans can or should depend on AI. Should humans supervise or approve certain AI-recommended outputs and decisions before they are implemented? Who is accountable for faults or hacking of AI-based technologies? There will be disputes over the inability of parties to fully comprehend or manage certain AI-powered operations, as well as disputes over what is fair in ADM.

For the efficiency and safety of AI-driven applications, judicial operators need to ensure that there is always a "human in the loop," i.e., AI never fully replaces humans so that adequately trained professionals validate AI decisions. AI is only as good as the data, human capital, and expertise of the interdisciplinary team involved in the development of the AI solution. An adequate AI and data governance framework should define the respective liabilities of all stakeholders, Judiciary stakeholders included. It should put in place the necessary conditions and guarantees to protect human rights while working towards the collective interest. This could be done by public certification of AI systems that would ensure that data and algorithm quality is guaranteed to

prevent deepening existing inequalities. Public certification of AI applications would build public trust and allow users to give informed consent.[46]

It is therefore important to be able to measure the level of risk and impact of different AI systems that might be deployed in the justice system. In this regard it is important to determine the requirement for human oversight, based on the use case, its sensitivity, the complexity and opacity of the algorithm, and the potential impact on human rights.[47] As an illustration, an AI chess player with low-risk might only necessitate a simple self-evaluation, user education, and in-house supervision. However, an AI surgeon with high-risk could mandate peer-reviewed evaluations, public records, significant human interventions, periodic training, and external scrutiny.

The Model Artificial Intelligence Governance Framework, Second Edition developed by the Government of Singapore (see Figure 8 below) outlines three broad approaches to human supervision of AI systems: (i) human-in-the-loop, (ii) human-out-of-the-loop, and (iii) human-on-the-loop. The extent to which human supervision is needed depends on the objectives of the AI system and a risk assessment, as illustrated by the examples below.

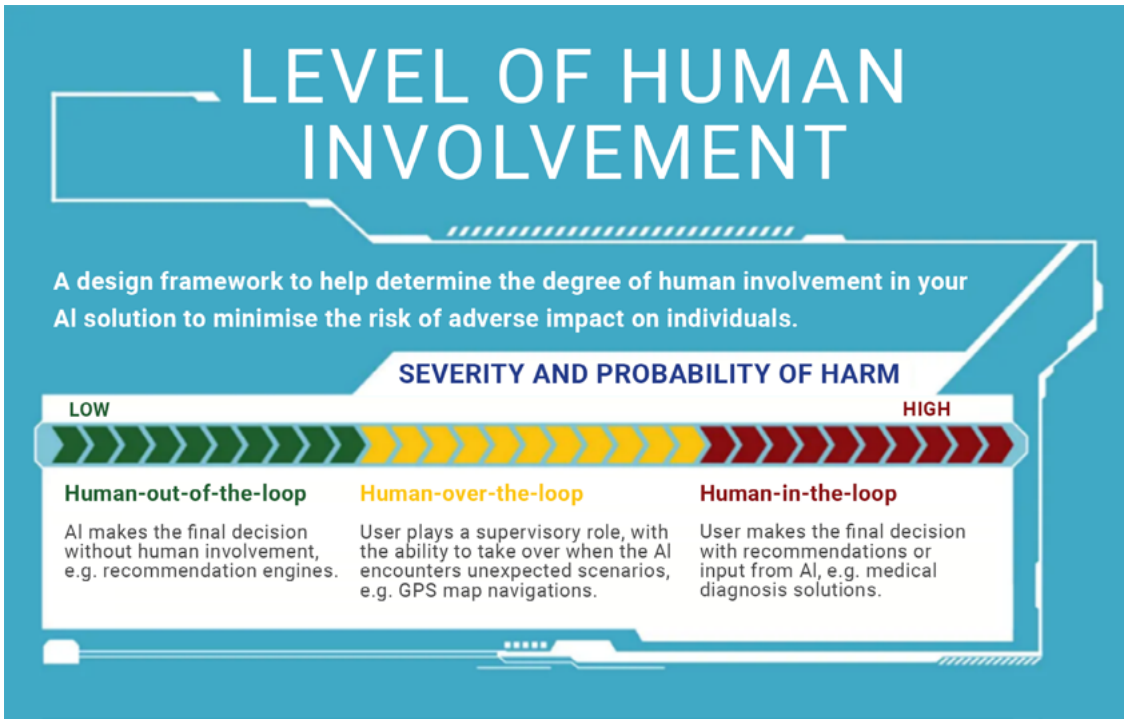
- **The term "Human-in-the-loop" (HITL)** refers to a process wherein an AI system is closely monitored by a human, who is responsible for making all final decisions. This is particularly important in fields like healthcare, where AI can provide invaluable support in making recommendations for cancer treatment, sepsis therapy, surgical planning, and more. While AI tools can help healthcare providers make informed decisions quickly and accurately, the ultimate responsibility for patient care always lies with the human expert.
- **The term "Human-out-of-the-loop"** pertains to the absence of human supervision in the decisions made by the AI system. This means that the AI system has complete control and there is no possibility for human intervention. An instance of this would be an AI-powered cybersecurity system that can detect and fix system vulnerabilities without the need for human involvement. Mayhem, the winning system in the Defense Advanced Research Projects Agency (DARPA) 2016 Cyber Grand Challenge, is an innovative system that constantly scans for any new vulnerabilities that could be exploited by hackers. When Mayhem detects a new bug, it automatically generates code to protect the software from this vulnerability. This system is an expert in prescriptive analytics, meaning it can detect and interact with machines without any human intervention. This is in contrast to traditional intrusion detection systems that rely on human input to anticipate cyber attacks.

46 Stankovich M. (2021). Regulating AI and Big Data Deployment in Healthcare: Proposing Robust and Sustainable Solutions for Developing Countries' Governments, available at: https://www.dai.com/uploads/regulating-ai-cda.pdf

47 According to the European Commission High-Level Expert Group on AI, 'HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impacts) and the ability to decide when and how to use the system in any particular situation', see: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1. See also the Model AI Governance Framework of Singapore, available at: https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework

- **The term "Human-on-the-loop"** refers to the involvement of humans in supervisory roles where they have the ability to take control when AI models encounter unexpected or undesirable situations. An effective way to understand this is through a GPS navigation system. The GPS system plans the route from point A to B and offers various options based on parameters such as shortest distance, shortest time, or avoiding toll roads. However, during navigation, the driver can still take over the GPS and modify the navigational parameters in the event of unexpected road congestion.

Figure 8. Level of human involvement in AI deployment



Source: IMDA, Singapore

It has to be noted that the HITL principle has its limitations due to automation bias discussed in Module 3, when humans are more predisposed to simple rubber stamp decisions made by algorithms especially in cases where there is a black box effect and humans might not be able to understand why this decision was taken.

5. Why is cybersecurity important in the context of AI?

Cybersecurity is the management of risks to the confidentiality, integrity, or availability of data and systems. It is an issue fundamental to any technology. AI processes/algorithms inherently process large datasets and frequently produce outputs with both virtual and tangible consequences. In addition to traditional threats, vulnerabilities unique to AI have been identified, including:

- Data poisoning during the training stage[48]
- Input attacks that manipulate data to alter the output[49]

Cyberattacks continue to increase in frequency, sophistication, and expense. In 2022, firms need an average of 207 days to detect a security incident and 70 days to contain it. As enterprises continue to quickly deploy technology across the value chainto deploy technology across the value chain quickly, the risk of business interruption assumes a central role. At home, embedded Internet of Things (IoT) devices continue to pose significant risks, and remote work introduces a complicated mix of vulnerabilities. Malicious actors can compromise AI systems to achieve various objectives, such as causing damage, evading detection, or degrading faith in a system.[50]

Malicious actors can compromise AI systems to achieve various objectives, such as causing damage, evading detection, or degrading faith in a system.[51]

Compared to traditional systems, AI-powered systems present unique features that can be vulnerable to cyber-attacks in non-traditional ways. For example, attackers may compromise a training dataset so that the resulting 'learning' of the system is not as intended. This type of attack is called data poisoning, and it takes advantage of AI's unique development process, which is the use of large size data. It is therefore important to provide for additional protection of AI systems. The rise of learning capabilities in AI technologies, such as deep learning and reinforcement learning has significant impact on cybersecurity and enables criminal actions more efficiently.[52] Therefore, protection of AI systems needs to be carefully considered and possible vulnerabilities identified to be able to put in place strong security measures to (a) guard against attacks, but also (b) detect attacks as soon as possible to mitigate against significant risks and harms.

48 Poremba S. (2021). Data Poisoning: When Attackers Turn AI and ML Against You, available at: https://securityintelligence.com/articles/data-poisoning-ai-and-machine-learning/

49 Comiter M. (2019). Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It, available at: https://www.belfercenter.org/publication/AttackingAI

50 Ibid.

51 Ibid.

52 Kaloudi N., Li J. (2020). The AI-Based Cyber Threat Landscape: A Survey. ACM Computing Surveys (CSUR), 53, 1–34, available at: https://www.researchgate.net/publication/339081899_The_AI-Based_Cyber_Threat_Landscape_A_Survey.

Cyber-attacks on AI systems occur in three different phases of AI development: 1) data preparation, 2) model training, and 3) model deployment: [53]

- During data preparation, attackers may target common data preparation components or libraries, or gain unauthorized access to data processing pipeline for tampering purposes.
- During the training phase, attackers can add, remove, or change training data (data poisoning). By doing this, attackers influence the resulting model.
- Attackers that have access to models can introduce changes to weights and algorithms in the model deployment stage (model tampering).[54]

Cybersecurity regulation

Cybersecurity regulation consists of directives that protect information technology and computer systems to compel private and public sector entities to protect their information systems and data from cyberattacks such as viruses, worms, Trojan horses, phishing, denial of service (DOS) attacks, unauthorized access (the theft of intellectual property or confidential information), and control system attacks[55].

Having this in mind, it is extremely important for judicial operators to take into consideration different cybersecurity laws and regulations and evaluate how AI can impact these regulations. For example, smart grids using AI systems, will significantly enhance the management of power consumption and distribution for the benefit of consumers, electricity providers, and grid operators. Nonetheless, enhanced operations and services will expose the entire energy network to new difficulties in communication and information system security. The vulnerabilities of communication networks and information systems could be exploited for financial or political reasons to cut power to broad areas or to launch cyberattacks against power producing units. AI can be used in misinformation and disinformation campaigns that could be used for Internet shutdowns and restricting access to information.[56] The box below describes the dangers associated with adversarial examples used by ML models.

53 Gartner (2020). Artificial Intelligence Under Attack: How to Identify and Mitigate Threats to Machine Learning, available at: https://www.gartner.com/en/documents/3989271; Wolff J. (2020). How to Improve Cybersecurity for Artificial Intelligence, available at: https://www.brookings.edu/articles/how-to-improve-cybersecurity-for-artificial-intelligence/

54 Ibid.

56 EU Cybersecurity Act (2019), available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R0881&qid=1694014957942. See also: https://web.archive.org/web/20100613183200/http://www.privacyrights.org/ar/ChronDataBreaches.htm

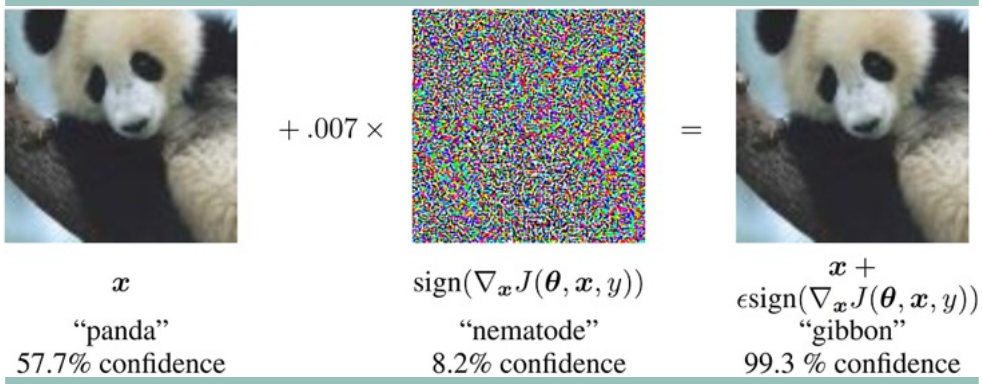
56 EU Cybersecurity Act (2019), available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R0881&qid=1694014957942. See also: https://web.archive.org/web/20100613183200/http://www.privacyrights.org/ar/ChronDataBreaches.htm

In depth: The dangers associated with adversarial examples used by ML models

Adversarial examples are inputs used by ML models that are purposefully generated by an attacker to make the model err while exhibiting a high level of confidence. Because many ML models, even cutting-edge neural networks[57], are susceptible to adversarial instances, this can pose a serious threat to AI safety and robustness.

Examples might be unnoticeable. The image of a panda below has undergone an undetectable small perturbation, or "adversarial input" inserted. It is intended to deceive the image-classification algorithm. This has resulted in the computer having a confidence level of 99.3% in classifying the panda as a gibbon.

Adversarial examples can be produced by printing an image on regular paper and taking a picture of it using a smartphone with a typical resolution. An antagonistic sticker on a stop sign could fool a self-driving car into thinking it is a "yield" sign or any other sign.[58]



Source: OECD, AI in society, available at: https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en

These AI systems' weaknesses against adversarial examples have detrimental effects on the security of AI systems. The adoption of critical systems like those used in autonomous transportation, medical imaging, and security and surveillance could potentially suffer seriously from the existence of cases where subtle but targeted perturbations lead models to be misled into gross miscalculation and incorrect decisions.

57 Neural networks are a type of ML technique that enables computers to learn how to perform tasks by analysing training examples. Typically, these examples are pre-labelled. For example, an object recognition system may receive thousands of labelled images of objects such as cars, houses, and coffee cups. Through analysis, it can identify patterns in the images that correspond with specific labels. A neural network is designed to loosely resemble the structure of the human brain, with thousands or millions of interconnected processing nodes. These nodes are typically organized into layers, and data flows through them in a single direction, making them "feed-forward". Each node receives data from nodes in the layer below it and sends data to nodes in the layer above it. Definition provided in Hardesty L. (2017). Explained neural networks, available at: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

58 Goodfellow I. J., Shlens J., Szegedy (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representation, available at: https://arxiv.org/pdf/1412.6572.pdf; Kurakin A., Goodfellow I., Bengio S. (2017). Adversarial examples in the physical world. ICLR Workshop, available at: https://arxiv.org/abs/1607.02533

6. Activities

These group activities are intended to encourage the training participants to discuss and debate various pertinent questions related to AI and its building blocks, and the risks associated with AI deployment in the Judiciary.

Activity 1 - Discussion time

Please discuss these questions with other training participants:

- How can a defendant legitimately contest the logic of an algorithm if the source code and (if applicable) training data or the datasets that will be required to reproduce the results are not made available to them?
- What information should be provided to the defendant to contest the logic of an algorithm?
- Is it sufficient for them to have access simply to the inputs and outputs generated by the algorithm?
- Should the defendant receive information on the margin of error of the algorithm(s) used?

Activity 2 - Discussion time

Please discuss these questions with other training participants:

- How can courts enforce due process of law if the algorithm deploys machine learning and no one, not even the developer, understands the ML "analysis" completely?
- How will courts assess the accuracy of algorithms, particularly when they forecast future human behaviour?

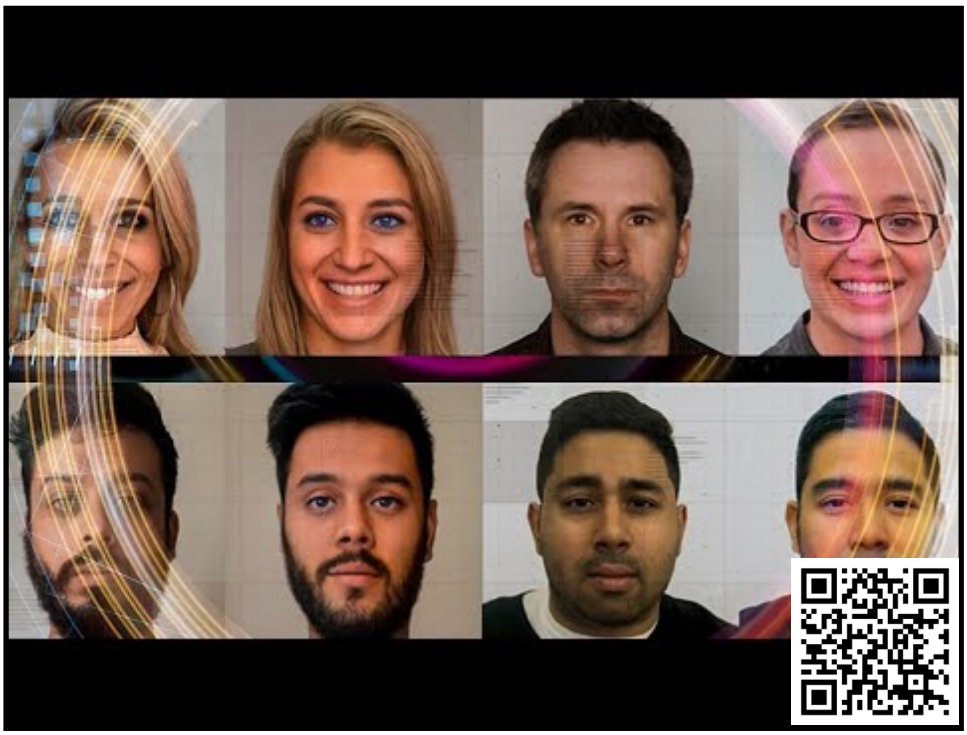
Activity 3 - Discussion time

Please discuss these questions with other training participants:

- What if the algorithms have been trained with past datasets that do not include the latest case law?
- What is the regime for the admissibility of evidence collected with the help of algorithms, especially by police investigators?
- Can this collection be considered irregular or unfair?
- Has the data been collected in compliance with data protection laws, and if that is not the case, how should the algorithm be treated?

Activity 4 - Discussion time

Training participants watch the video and discuss different societal impacts of AI bias.



Source: BBC, https://youtu.be/b4UyT85H3Hg

Activity 5 - Training participants discuss the following issues related to the application of AI in judicial operations.

Often, AI models cannot provide human-comprehensible justifications for their decisions or recommendations. Many AI algorithms "learn on their own", i.e., self-learning ML (Also, read and refer to the human in the loop principle in Module 4. Try answering the following questions while discussing with other training participants:

- How does your capacity to comprehend or probe an AI model's output affect its evidentiary value in litigation proceedings?
- What legal and social responsibilities should we give to algorithms shielded behind statistically data-derived 'impartiality'?
- Who is liable when AI gets it wrong?

There is much debate as to who amongst the various players and actors across the design, development and deployment lifecycle of AI and autonomous systems should be responsible and liable to account for any damages that might be caused. A complex AI eco-system and the multiplicity of actors make it difficult to determine who may be held liable for the damage caused to the claimant(s), as the damage may result from a series of intertwined causes by multiple actors.

- Would autonomy and self-learning capabilities alter the chain of responsibility of the producer or developer as the "AI-driven or otherwise automated machine which, after consideration of certain data, has evolved over time through its self-learning abilities enabled by ML and/

or deep learning techniques taken an autonomous decision and caused harm to a human's life, health or property"?

- How will the capabilities of unsupervised ML systems affect issues of liability. For instance, a challenge is the reliance on external data – where such data is supplied from external sources, proving both defectiveness and a causal link with the injury or damage sustained could be very difficult.
- Does "inserting a layer of inscrutable, unintuitive, and statistically derived code in between a human decision maker and the consequences of that decision, AI disrupts our typical understanding of responsibility for choices gone wrong"? Or should the producer or programmer foresee the potential loss or damage even when it may be difficult to anticipate—particularly in unusual circumstances, the actions of an autonomous system? These questions will become more critical as more and more autonomous decisions are made by AI systems.
- Which levels of uncertainty in ML model outputs will the courts accept, and under what conditions? How do various levels of ML model certainty relate to various evidence standards? (i.e., when does X, Y, or Z degree of correlation [minus X%, Y%, or Z% of uncertainty] equal some legal standard of proof (such as "clear and convincing," "preponderance of evidence," or "beyond a reasonable doubt)?
- Will this be an issue left to the discretion of individual courts, or judges? Or will national or regional standards be developed and implemented? Should these requirements be rigid or flexible?
- Most ML processes are iterative and self-learn as they adjust formulae and accuracy by processing new data. Do you think that such ML applications, as they constantly change, might need to be re-litigated on an ongoing basis, if used as evidence?

Source: AAAS, Artificial Intelligence and the Courts: Materials for Judges, available at: https://www.aaas.org/ai2/projects/law/judicialpapers.

7. Resources

1. AAAS, Artificial Intelligence and the Courts: Materials for Judges, available at: https://www.aaas.org/ai2/projects/law/judicialpapers
2. AccessNow (2018). Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems, available at: https://www.accessnow.org/press-release/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/
3. Amnesty International (2017). Artificial Intelligence for Good, available at: https://www.amnesty.org/en/latest/news/2017/06/artificial-intelligence-for-good/
4. Amnesty International (2021). Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal, available at: Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal - Amnesty International
5. Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, available at: https://ssrn.com/abstract=4162985
6. Buolamwini J., Gebru T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 81, 77–91, available at: https://proceedings.mlr.press/v81/buolamwini18a.html
7. Burgess M. (2023). The Security Hole at the Heart of ChatGPT and Bing. available at: https://www.wired.co.uk/article/chatgpt-prompt-injection-attack-security
8. Burrell J. (2015). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms available at: https://ssrn.com/abstract=2660674 or http://dx.doi.org/10.2139/ssrn.2660674
9. Conn A. (2017). Artificial Intelligence: The Challenge to Keep It Safe., available at: https://futureoflife.org/ai/safety-principle/European Union Agency for Fundamental Rights (2019), available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf
10. IEEE (2019). Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. First Edition, available at: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf
11. International Commissioner's Office, Explaining decision made with AI, available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/
12. McGregor L., Murray D., Ng V. (2019). International Human Rights Law as a Framework for Algorithmic Accountability, International & Comparative Law Quarterly, 68(2), 309–343, available at: https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/international-human-rights-law-as-a-framework-for-algorithmic-accountability/1D6D0A456B36BA7512A6AFF17F16E9B6
13. National Science and Technology Council: Committee on Technology (2016). Preparing for the Future of Artificial Intelligence. Washington, D.C.: Executive Office of the President, 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

14. Noble S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism, New York University Press.

15. Obermeyer Z., Powers B., Vogeli C., Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations, Science, 366(6464), 447–453, available at: https://www.science.org/doi/10.1126/science.aax2342

16. OECD (2022). Framework for the Classification of AI systems, available at: https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html

17. O'Neil C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York: Crown.

18. Stanford (2022). Artificial Intelligence Index Report, available at: 2022-AI-Index-Report_Master.pdf (stanford.edu)

19. The Alan Turing Institute, Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A proposal prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence, available at: https://www.turing.ac.uk/news/publications/ai-human-rights-democracy-and-rule-law-primer-prepared-council-europe

20. The Royal Society (2012). Machine Learning: The Power and Promise of Computers that Learn by Example, available at: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf%2016

21. UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

22. Ward J. (2019). 10 Things Judges Should Know About AI, Judicature, 103(1), available at: https://judicature.duke.edu/articles/10-things-judges-should-know-about-ai.

23. Weinberger D. (2017). Our Machines Now Have Knowledge We'll Never Understand, available at: https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/

24. Wong A. (2020). The Laws and Regulation of AI and Autonomous Systems. In: Strous L., Johnson R., Grier D. A., Swade D. (eds) Unimagined Futures – ICT Opportunities and Challenges, IFIP Advances in Information and Communication Technology(), 555, available at: https://link.springer.com/chapter/10.1007/978-3-030-64246-4_4

25. Wong A. (2021). Ethics and Regulation of Artificial Intelligence. In: Mercier-Laurent E., Kayalica M.Ö., Owoc M.L. (eds) Artificial Intelligence for Knowledge Management, AI4KM, IFIP Advances in Information and Communication Technology, 614, available at: https://www.researchgate.net/publication/352477342_Ethics_and_Regulation_of_Artificial_Intelligence

26. Wong A. (2023). Generative AI: The Global debate and controversies on use of copyrighted content as training data, available at: https://unctad.org/news/cstd-dialogue-anthony-wong

Module 2

AI Adoption in the Judiciary

Module two discusses AI adoption in the Judiciary. It presents the different applications of AI in the Judiciary, such as e-discovery and document review, use of generative AI to assist in the drafting of documents, predictive analytics, risk assessment tools, dispute resolution, language recognition, digital file and case management. The Module then highlights case studies on AI deployment in the Judiciary, discussing some of the opportunities and challenges encountered by judicial systems worldwide in the use of AI.

What will you learn?

After completing this module, the participants will be able to:

- Understand the different applications of AI in the Judiciary;
- Understand the challenges and opportunities related to the deployment of AI systems in the Judiciary through the case studies presented in the module.

1. What are the applications of AI in the Judiciary?

Lawyers, law firms, courts, and government agencies are using AI for different purposes. For instance, lawyers are using AI for legal research and to find relevant precedents to strengthen their arguments. Law firms are using it to forecast case outcomes, assess success chances, and counsel clients regarding legal proceedings. AI has also been used by lawyers to forecast how particular judges would rule on various topics. Similarly, government entities are using AI assess the likelihood success in pursuing particular courses of action against individuals and businesses, such as in tax-related cases.

In Buenos Aires, Argentina, the tax prosecutors use AI systems to write court rulings[59]. The Hangzhou Internet Court has implemented an evidence analysis system that uses cutting-edge technologies such as blockchain, AI, big-data, and cloud computing. This system analyzes and compares all evidence presented by both parties, transforming it into a list of evidence and relevant exhibits. The information is then sorted and classified before being visually presented to the human judge for their consideration.[60] In Mexico, courts can use AI to give advice on determining whether someone is entitled to a form of social security or not. A program named Expertius grounds its calculations on information about past claims, results of the claims, hearing records, and final judgments.[61]

Another example is the Colombian justice system, which is exploring ways to reduce the workload of human judges. The Colombian Constitutional Court is currently developing an AI system called PretorIA to assist in the selection of legal guardians. PretorIA does not replace humans in this process, but rather streamlines the task by analyzing guardianship sentences and providing more refined information to those responsible for identifying individuals who can be selected as guardians.[62]

The push for efficient justice amidst budget constraints

As with other consumer services, courts are expected to provide modern, digital, and responsive judicial services, while reducing the pendency of cases in a context of increasing budgetary constraints. AI-enabled justice systems promise to scale up quality of services while reducing expenses related to judicial operations.[63]

59 Dejusticia (2021). Conoce nuestra Investigación sobre PretorIA, la tecnología que incorpora la Inteligencia Artificial a la Corte Constitucional, available at: https://www.dejusticia.org/conoce-nuestra-investigacion-sobre-pretoria-la-tecnologia-que-incorpora-la-inteligencia-artificial-a-la-corte-constitucional/

60 Xuan H. (2021). One-Click Access to Evidence Analysis Results. Hangzhou Internet Court Launches Intelligent Evidence Analysis System, China Courts Network, available at: https://www.chinacourt.org/article/detail/2019/12/id/4747683.shtml

61 Goretty C., Martinez B. (2012). La inteligencia artificial y su aplicación al campo del Derecho, Alegatos, 82, 827–846, available at: http://alegatos.azc.uam.mx/index.php/ra/article/viewFile/205/184

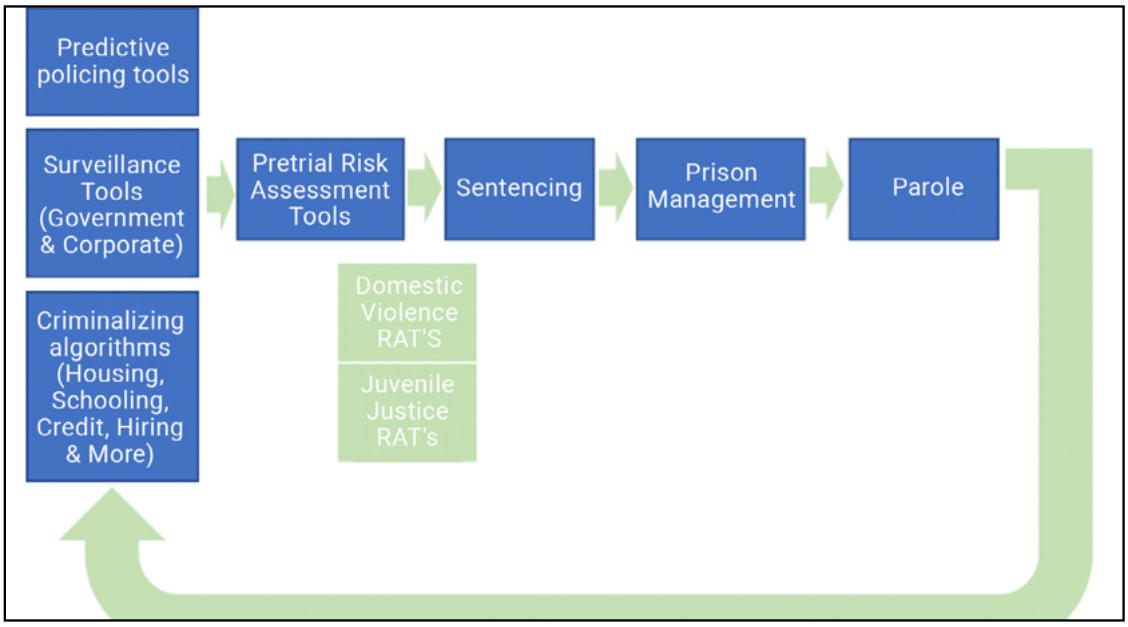
62 https://www.dejusticia.org/conoce-nuestra-investigacion-sobre-pretoria-la-tecnologia-que-incorpora-la-inteligencia-artificial-a-la-corte-constitucional/

63 Wu J. (2019). AI Goes to Court: The Growing Landscape of AI for Access to Justice, available at: https://medium.com/legal-design-and-innovation/ai-goes-to-court-the-growing-landscape-of-ai-for-access-to-justice-3f58aca4306f

When deployed with human rights and ethical safeguards, AI systems can make legal procedures more accessible to a wider group of individuals, in multiple languages, and at lower costs. For instance, estimates show that using ML in e-discovery by presenting the documents in conceptual clusters can increase the review speed by 15 to 20 per cent. This is a significant cost saver.[64]

On the other hand, AI development and deployment in judicial operations can impact fundamental rights. AI technologies contain embedded bias (discussed in Module 3), and they are oftentimes black boxes- (discussed in Module 1). Therefore, the rule of law and the preservation of human rights must continue to be at the forefront of administration of justice.[65]

Figure 9. A simplistic cycle of algorithmic use in criminal justice



Source: EPIC, AI in the criminal justice system, available at: https://epic.org/issues/ai/ai-in-the-criminal-justice-system/

Digitization of Court documents is an essential first step towards AI use

Digitization of court documents has enabled courts and other judicial operators to rely on AI assistance for administrative functions. AI algorithms are increasingly being used in the context of the civil and criminal justice systems to support human decision-making.[66] AI systems are tested to identify patterns in complex judicial decision-making and predict decision outcomes. As AI systems gather and analyse vast troves of information, identify patterns, predict optimal approaches, detect anomalies, classify issues, and draft documents, the promise is that

64 Deloitte, Artificial intelligence and machine learning in e-discovery and beyond: Driving efficiencies in e-discovery using AI, available at: https://www2.deloitte.com/ch/en/pages/forensics/articles/AI-and-machine-learning-in-E-discovery.html,

65 On the impact of AI on human rights when applied in the judicial systems, see also UNESCO (2021). Global Toolkit for Judicial Actors: International legal standards on freedom of expression, access to information and safety of journalists, Module 5, p. 164, available at: https://unesdoc.unesco.org/ark:/48223/pf0000378755

66 European Parliament (2019). A governance framework for algorithmic accountability and transparency, available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

court systems will become more efficient and be able to prioritise time and resources to ensure timely justice.

In the criminal justice system, AI models have been deployed to monitor and recognize defendants; support sentencing and bail decisions; and support assessment of evidence.[67] Figure 9 gives a simple overview of AI algorithmic use in the criminal justice system.

In the civil justice system, AI has been deployed in family, housing, debt, employment, and consumer litigation.[68] Civil courts are increasingly collecting data about administration, pleadings, litigant behaviour, and decisions. This offers opportunities for automating certain judicial functions, such as docket management, scheduling hearings and trials, and managing jury functions, which in turn can lead to greater efficiency.[69] For example, AI is used to pre-draft judgment templates for judges, make predictions or sentencing recommendations for bail, sentencing and financial calculations. It is also used to assess the outcome of cases based on the past activities of prosecutors and judges. An AI tool can provide information to a judge that factors in a wide amount of case law and can decrease the research time in the preparation of decisions.

Using an AI algorithm created by researchers at Université Catholique of Leuven (UCL), the University of Sheffield, and the University of Pennsylvania, the European Court of Human Rights judicial rulings have been anticipated with an accuracy of 79%.[70] Dr Nikolaos Aletras, who led the study at UCL Computer Science explained that "We don't see AI replacing judges or lawyers, but we think they'd find it useful for rapidly identifying patterns in cases that lead to certain outcomes. It could also be a valuable tool for highlighting which cases are most likely to be violations of the European Convention on Human Rights".[71]

The challenge of AI systems being perceived as more objective than humans

However, given the high caseloads and lack of adequate resources that plague most judicial systems, there is a risk that judges will improperly use AI-based support systems to "delegate" decisions to technological systems that were not designed for that purpose but are perceived as more objective than they are. In order not to jeopardize the right to a fair trial, great care should be taken to evaluate what such devices are capable of and under what conditions they may be deployed. This is especially true when such systems are used in delivering parole decisions. In an algorithm driven justice system, judges should not be the mere appliers of algorithms, but also their critical evaluators. The table below outlines the key positive and negative implications of using ADM and AI in the justice system.

67 Završnik A. (2020). Criminal justice, artificial intelligence systems, and human rights. ERA Forum. 20, 567-583, available at: https://link.springer.com/article/10.1007/s12027-020-00602-0

68 Cabral J. E, Chavan A., Clarke T. M., Greacen J., Hough B. R., Rexer L., Ribadeneyra L., Zorza R. (2012). Using Technology to enhance access to justice, available at: http://jolt.law.harvard.edu/articles/pdf/v26/26HarvJLTech241.pdf

69 Martin A. (2010). Automated Debt-Collection Lawsuits Engulf Courts, available at: https://www.nytimes.com/2010/07/13/business/13collection.html

70 UCL (2016). AI predicts outcomes of human rights trials, available at: https://www.ucl.ac.uk/news/2016/oct/ai-predicts-outcomes-human-rights-trials

71 Ibid.

Table 3. Positive and negative implications of use of AI in the justice system

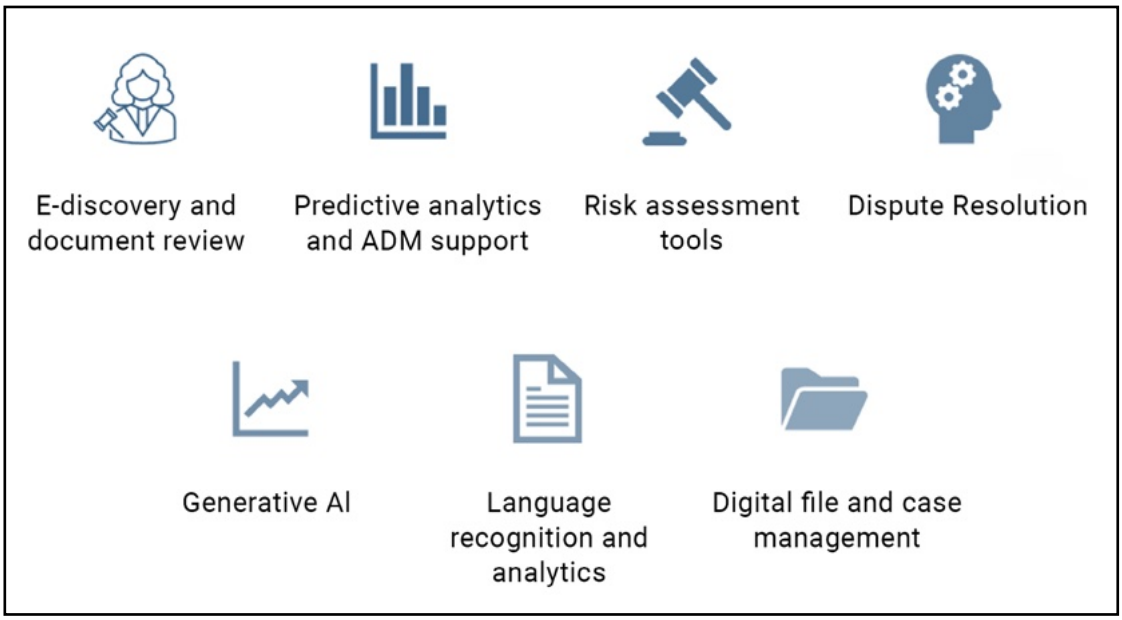
	Positive implications	Negative implications
Judicial excellence	Gives judges a quick analysis of range of cases and factors; Speeds up research and drafting; Process optimisation, cost reduction, increased agility, productivity gains, elimination of mechanical and repetitive work increase legal security	Embedded racial, gender/sex and other types of biases; Reduces judicial discretion and human element in decision making; Reduces judicial discretion and human element in decision making; Complicated to use; Threat to judicial independence, automated bias; Profiling of the judges can affect the fundamental right to the protection of personal of personal data, may create pressures and affect judicial independence
Privacy and security	Automatic security protocols and data cleansing, which allows for higher accuracy in AI outputs	Hacking, data breaches
Data ownership	Aggregated data by AI systems can be used to identify trends, service gaps and innovation[72].	Depending on the ownership of the system private sector partners might have access to personal data; Aggregated data can be used to target and discriminate against individuals or groups; Limited regulation of data ownership limits the protection of rights and redress for people impacted by AI systems
Rule of law	Prevents powerful interest from capturing the justice system	Can encroach on fundamental rights as discussed in Module 4 Threats to democracy such as disinformation, misinformation, hoaxes, propaganda, deep fake, influence operations or manipulation of public opinion, mainly in electoral processes.
Access to justice	Can identify patterns of bias against vulnerable groups in decision making and services Can make court timelines faster and more predictable	Not uniformly available to parties to analyze data or support their case due to infrastructural and access issues (electricity, internet, hardware) The lack of training from judicial operators and assistants might impact the positive outcomes AI could bring.

Adapted from UNDP (2021) Emerging Technologies and Judicial Integrity Toolkit for Judges.

Source: https://www.undp.org/asia-pacific/emerging-technologies-and-judicial-integrity

72 IBM (2021). Data aggregation involves gathering a significant amount of information from a database and presenting it in a more manageable and inclusive format, available at: https://www.ibm.com/docs/en/tnpm/1.4.2?topic=data-aggregation

Figure 10. Key AI applications in the Judiciary



Source: Authors.

E-discovery and document review

AI tools are used in the Judiciary to identify, sort through and review (i) legal rules, legal holdings and factual findings; (ii) arguments explaining conclusions and explanations of reasons, and (iii) specific legal considerations and evidentiary elements.

E-discovery is the identification, collection, and production of electronically stored information (ESI) in response to a disclosure request in a judicial proceeding or investigation. ESI can consist of emails, documents, presentations, databases, audio and video files, and websites.[73]

Activity: Think about how AI can change the discovery process and discuss it with other training participants.

Issues to consider: What will be the standards for admissibility of statements or other evidence, or insights generated by AI and/or relied upon (or rejected) by humans? How will we assess its credibility or authenticity?

73 https://cdslegal.com/knowledge/the-basics-what-is-e-discovery/

E-discovery relies on clustering, an example of unsupervised ML, where "similar" items (e.g., documents) are grouped together so that users can recognize their similar characteristics and learn about the composition of the data set. Users have no control over the dimension(s) along which "similarity" is defined and do not have to label examples of items in each cluster to train the system. However, the designer of the system must specify the features along which item similarity is to be measured and the number of clusters.[74] For instance, if the ML system is instructed to identify any information about tennis and baseball in the files, the algorithm will also cluster files that contain information about all kinds of sports.[75] Similarly, a search for "little brown envelope" or "grease" will cluster information about everything related to corruption.[76]

Concept search is another unsupervised ML method used in e-discovery, where the computer learns the context in which words are used and models the relationships among words. Users can then search by meaning and not by individual terms. It is likely that a document containing words such as "lawyer," "contract," or "civil litigation" is a legal document. The use of any of these words can lead to the conclusion that the topic of the document is legal.[77]

Technology Assisted Review (TAR) or predictive coding is a supervised ML technique in which computers learn to distinguish relevant from irrelevant documents based on the coding done by human reviewers, and then classify unlabelled documents without assistance.[78] For instance, CLAUDETTE (Automated CLAUse DETectEr) is an interdisciplinary research project hosted at the Law Department of the European University Institute and a platform for analysis and automated annotation of legal documents, and anomaly detection.[79]

Further, AI tools can be deployed for anonymizing personal, confidential, or privileged information included in electronic records. This can aid compliance with data protection regulations.[80]

Activity: How does CLAUDETTE function?

CLAUDETTE's goal is to empower consumers and civil society by ultimately developing user-end tools that enable everyone to easily evaluate the fairness of consumer contracts and privacy regulations before utilizing internet platforms. The technology is currently in the experimental, laboratory stage – and training participants can access it here: http://claudette.eui.eu/demo

Training participants watch the video (http://claudette.eui.eu/claudette.mp4) and discuss if they have similar platforms in their respective jurisdictions. What are the advantages and disadvantages of TAR?

74 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, disponible en: https://edrm.net/download/152621/75 IBM (2021). Data aggregation involves gathering a significant amount of information from a database and presenting it in a more manageable
75 Deloitte. Artificial intelligence and machine learning in e-discovery and beyondnDriving efficiencies in e-discovery using AI, available at: https://www2.deloitte.com/ch/en/pages/forensics/articles/AI-and-machine-learning-in-E-discovery.html
76 Ibid.
77 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, available at: https://edrm.net/download/152621/
78 Ibid.
79 EUI. CLAUDETTE, available at: http://claudette.eui.eu/about/index.html
80 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, available at: https://edrm.net/download/152621/

Predictive analytics and ADM support

Frequently, AI systems are used as forecasting tools. They analyze big quantities of data, including historical data, to assess risks and predict future trends using algorithms. Training data may contain, criminal records, arrest records, crime statistics, records of police interventions in certain neighbourhoods, social media posts, communications data, and travel records. Predictive systems can assist judges in having better awareness of trends in the case law and in anticipating how a possible decision will stand in the context of the case law.[81]

Predictive analytics is the umbrella category of statistical tools and models, e.g., ML systems, that use and analyze historical data to create predictions about the future to guide decision making. These predictions can be low risk (e.g., which movie to recommend), medium risk (which loan application to propose accepting), or high risk (which defendant is most likely to engage in a particular behaviour) .[82]

The development of AI applications that forecast how a court will determine a claim, case, or settlement is a fast growing application of AI in the justice sector. For instance, AI technologies are already being used in profiling people, identifying places as likely sites of criminal activity, or flagging future reoffenders.[83] These practices are highly controversial, as elaborated in Modules 3 and 4.

One such example is the EXPERTIUS system in Mexico, which advises judges and clerks on whether a plaintiff is eligible for a pension. The program consists of three modules; first, it provides judges and clerks with an opportunity to understand the process (the tutorial module); second, it allows users to provide evidence in support of their case and assign 'weights' to each piece of supporting documentation (the inferential module); and third, it enables users to calculate the amount of pension to which they are entitled based on specified socio-economic criteria (the financial module).[84]

81 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, available at: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

82 AAAS, Artificial Intelligence and the Courts: Materials for Judges, available at: https://www.aaas.org/ai2/projects/law/judicialpapers

83 RAND (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operation, available at: https://www.rand.org/pubs/research_reports/RR233.html

84 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, available at: https://ssrn.com/abstract=4162985

Case Study: The case of the Australian Split UP system

A group of AI experts and attorneys have developed the Split-Up system that is used in Australian Family Law courts. The Split-Up system uses rules-based reasoning in conjunction with neural networks to anticipate the outcomes of property disputes in divorce and other family law matters.

The Split-Up system is used by judges to support their decision-making by assisting them in identifying the marital assets that should be included in a settlement. The system helps the judge determine what percentage of the common pool each party should receive based on factors such as contributions, income sources, and future needs. The system analyses 94 key elements using statistical techniques based on neural network architecture. The judge can then propose a final property order based on the analysis performed by the algorithm. The system also aims to provide clear justifications for its decisions.

One challenge in terms of bias when using systems such as Split-Up is that the data used in this context (divorce disputes are usually marked by gender imbalances and historic data can present a pattern of discrimination) might be read as ground-truth by machines. Judicial operators should be made aware of these challenges and risks that come with AI systems such as Split-Up.

Source: Zeleznikow J., Stranieri A. (1995). The split-up system: integrating neural networks and rule-based reasoning in the legal domain, ICAIL '95: Proceedings of the 5th international conference on Artificial intelligence and law, 185–194, available at: https://dl.acm.org/doi/10.1145/222092.222235

Risk assessment tools (risk prediction, risk modelling and social scoring)

Increasingly, data-driven risk assessment tools are used to anticipate the probability of future criminal behaviour. In several countries, these technologies are being used to aid decision-making in the criminal justice system, including judgments regarding sentencing, bail, and post-sentence limitations for those deemed likely to commit other crimes. These tools leverage historical data to assess the likelihood of an individual being a "high," "medium," or "low" risk for missing their court dates or getting re-arrested. The algorithm considers factors such as criminal record and age at the time of arrest, and generates a score that judges use to decide whether to hold someone in jail or release them.[85]

To assess a person's risk of reoffending and identify intervention areas, risk assessment tools are used at various phases of the legal process. For instance, risk assessments are used:

i) Before trial to guide choices on release awaiting resolution or incarceration.

85 Wykstra S. (2018). Bail reform, which could save millions of unconvicted people from jail, explained, available at: https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality

ii) By probation and parole departments to determine the appropriate amount of supervision, which may include electronic monitoring and home confinement.

iii) As part of re-entry and supervision plans, case managers and treatment providers deploy risk assessments to pinpoint client needs and connect them to the right services.[86]

Risk assessment techniques, according to their proponents, make the criminal justice system more equitable.[87] The proponents of such systems argue that AI could substitute judges' intuition and bias, particularly racial bias, with a risk assessment score that appears to be more "objective."[88]

However, in practice, numerous studies have shown that these tools might embed and amplify biases towards marginalized and vulnerable populations. Several human rights can be implicated using AI in the criminal justice system, including the rights to equality and non-discrimination, equality before the law, personal security and liberty, the right to privacy, the right to a fair and public hearing, procedural fairness, and the presumption of innocence (see Figure 11 that gives an overview of how criminal justice risk assessment tools impact human rights; for specific examples please refer to Module 4 of this Toolkit).[89] To illustrate these points, some risk assessment tools rely on data from police calls, which can be an unreliable indicator of actual crime patterns (vis-à-vis arrest records). This data is often further distorted by racial biases, as seen in the infamous case of Amy Cooper, who called the police on a Black bird-watcher for simply asking her to leash her dog in Central Park.[90] It is crucial to understand that just because a call is made to report a crime, it does not necessarily mean that a crime has actually occurred. However, such calls can be used as data points in risk assessment systems to justify dispatching police to a particular neighborhood or even targeting a specific individual, thus creating a feedback loop where data-driven technologies legitimize discriminatory policing.[91]

In the case of Ewert v. Canada, the Supreme Court of Canada emphasized that risk assessment tools that are created and verified using data from majority groups may not be accurate in predicting the same features in minority groups.[92]

86 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, available at: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

87 Hao K., Stray J. (2019). Can you make AI fairer than a judge? Play our courtroom algorithm game, available at: https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

88 Wykstra S. (2018). Bail reform, which could save millions of unconvicted people from jail, explained, available at: https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality

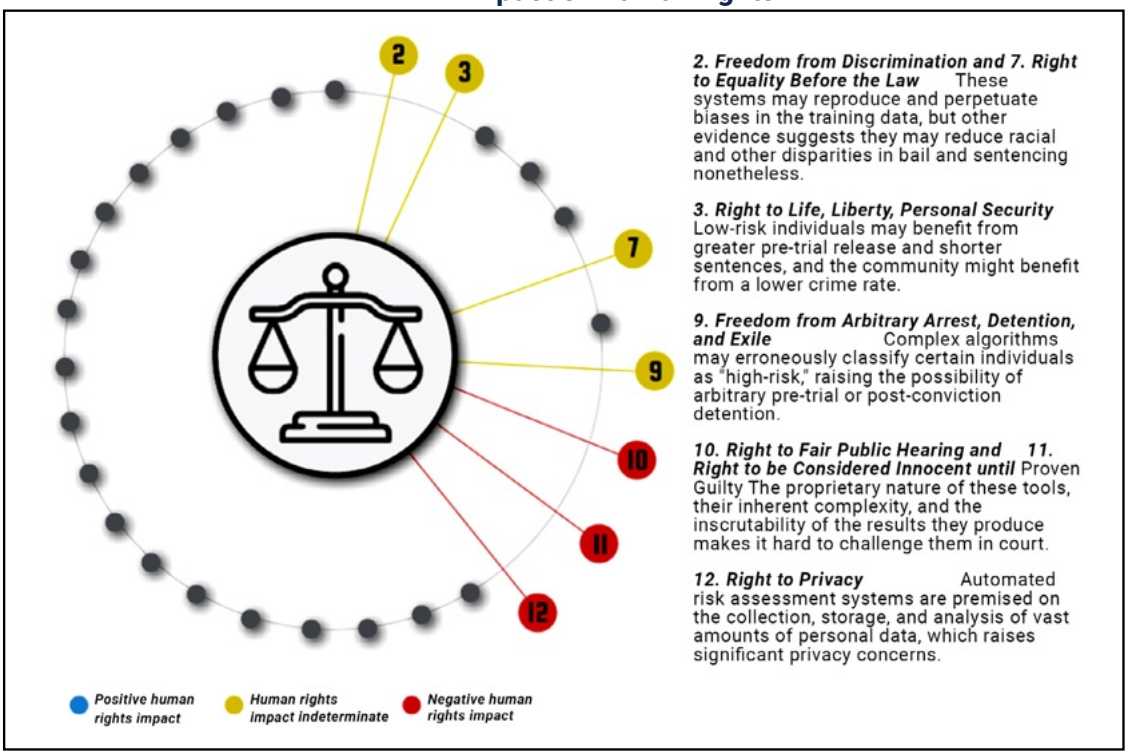
89 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, available at: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

90 Nir S. M. (2020). How 2 Lives Collided in Central Park, Rattling the Nation, available at: https://www.nytimes.com/2020/06/14/nyregion/central-park-amy-cooper-christian-racism.html

91 Heaven W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled, available at: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

92 Available at: https://www.scc-csc.ca/judgments-jugements/cb/2018/37233/

Figure 11. Criminal justice risk assessment tools' impact on human rights



Source: Raso F., Hilligoss H., Krishnamurthy V., Bavitz C., Kim L. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks, available at: https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights

Dispute resolution

AI systems can be used to forecast how a case will be decided, thereby providing plaintiffs with a better grasp of their options, or generating a settlement proposal. In this approach, judicial decision prediction could facilitate access to justice. Such systems may be integrated into online court platforms where individuals explore their legal alternatives or enter and exchange case-related information. The AI system would assist litigants in making better filing decisions, and it would assist courts in accelerating decision-making by supplementing or replacing findings by judges.[93]

Many Online Dispute Resolution (ODR) platforms do not use AI, but instead serve as a platform for litigants' job coordination and simplification. However, ODR platforms such as Rechtwijzer, used in the Netherlands[94], MyLaw BC, Canada[95], and the ODR used by the British Columbia Civil Resolution Tribunal (CRT), Canada[96], use AI systems to determine which parties can use the platform to resolve a dispute, as well as to automate decision-making and settlement or outcome recommendation.

For instance, the British Columbia CRT dispute resolution procedure begins with the Solution Explorer, an AI expert system, which employs a question-and-answer structure to provide users with individualized, simple

93 Wu J. (2019). AI Goes to Court: The Growing Landscape of AI for Access to Justice, available at: https://medium.com/legal-design-and-innovation/ai-goes-to-court-the-growing-landscape-of-ai-for-access-to-justice-3f58aca4306f
94 See: https://rechtwijzer.nl/
95 See: https://family.legalaid.bc.ca/retiring-mylawbc
96 See: https://civilresolutionbc.ca/

language legal information and free self-help resources to settle their problem without the need to submit a CRT claim. Lawyers from across British Columbia contributed to produce legal content for the Solution Explorer. Knowledge engineers visited attorneys and interviewed them about the most frequent problems seen in their practice areas as well as the legal facts they believe the public should be aware of. The CRT team then organized this data into extensive mind maps, making sure the language and content are clear and aimed at school Grade 6 readers.[97]

Activity: The example of the British Columbia CRT Solution Explorer

Training participants watch the video below and discuss if similar solutions that use AI expert systems can be found in their jurisdictions.

Source: https://youtu.be/ueVUETHy8gc

Case Study: Jury bot

Every year, the Superior Court of Los Angeles County deals with about 1.2 million new traffic citations. Several years ago, people had to wait as long as 2.5 hours to see a clerk for their traffic problem because of a state financial crisis that resulted in courthouse closures and reduced personnel.[98] Now, an online assistant for the Superior Court of Los Angeles, assists people with their traffic tickets. The jury bot uses ML translation services, and natural language understanding. It assists more than 5,000 citizens each week and speaks five languages.

Source: The Superior Court of California, County of Los Angeles, available at: https://ww2.lacourt.org/traffic/ui/trafficOS.aspx?s=1&language=2

97 Salter S. (2018). What is the Solution Explorer?, available at: https://www.cbabc.org/BarTalk/Articles/2018/April/Features/What-is-the-Solution-Explorer

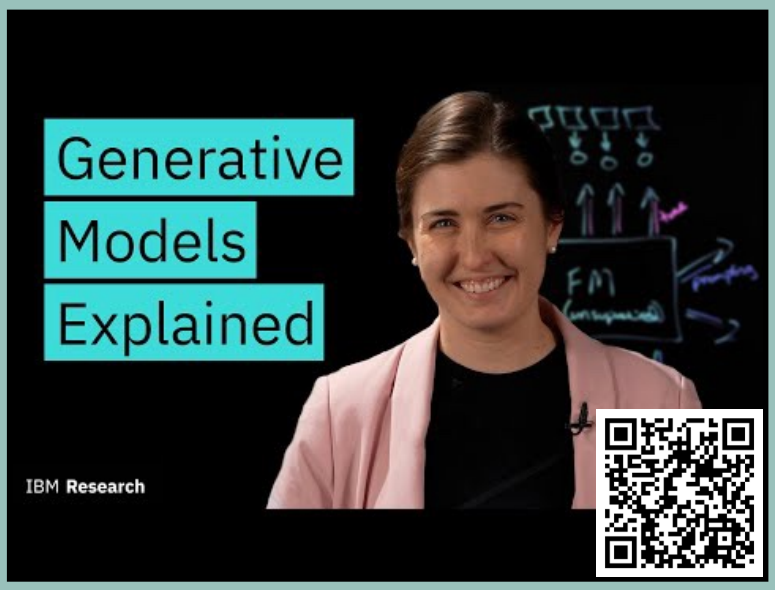
98 SRLN (2023). News: Gina — LA›s Online Traffic Avatar Radically Changes Customer Experience (Los Angeles 2016), available at: https://www.srln.org/node/1186/gina-las-online-traffic-avatar-radically-changes-customer-experience-news-2016

In Australia, the state of Victoria is piloting ODR platforms through its VCAT pilot for small claims.[99] These pilots use platforms such as Modria, Modron, and Matterhorn by Court Innovations. It is unclear to what extent AI is included into these systems, but they appear to be mostly platforms for logging facts and preferences, interaction between parties, and drafting/signing agreements (without any algorithm or AI tool deciding or crafting a strategy for parties). If the pilots are successful and become ongoing initiatives, future iterations may include additional AI-powered recommendations or decision aids.[100]

Generative AI

The field of generative AI is currently experiencing an era of unprecedented progress. These machine learning algorithms have been designed to create new content, including audio, code, images, text, simulations, and videos. Recently, chatbots such as ChatGPT, Bard, and Copilot have been developed that use large language models (LLMs) to perform various functions, such as research gathering, legal case file compilation, repetitive clerical task automation, and online search. This innovative technology has the potential to significantly increase efficiency and productivity by simplifying specific processes and decisions, such as streamlining note processing or helping educators teach critical thinking skills.[101]

Discussion point: Training participants watch the video and discuss how Generative AI has influenced their lives. Have they tried using it in decision making processes? What are the key opportunities and challenges related to Generative AI?



Source: https://www.youtube.com/watch?v=hfIUstzHs9A

99 Legaltech News (2020). A Future ODR Roadmap for Courts Post-COVID-19, available at: https://www.law.com/legaltechnews/2020/06/23/a-future-odr-roadmap-for-courts-post-covid-19/
100 Ibid.
101 Routley N. (2023). What is generative AI? An AI explains, available at: https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/.

Generative AI systems can generate text, including legal arguments or research, by predicting the appropriate text to follow a given input using patterns learned from extensive data sets. This makes generative AI a potent tool in several fields, including the legal profession. While some generative AI tools operate within a closed universe of information, others are open and have wider data access, such as through web plugins or internet connections.[102]

Many governments around the globe have started curtailing the use of large language models (LLMs)[103]. EU's AI Act also contains rules for general-purpose AI, or AI systems that may be deployed for a variety of tasks with various levels of risk. Similar technologies include ChatGPT and other LLM generative AI systems. In another example, due to data protection and privacy concerns, the Italian data protection regulator issued a temporary ban on ChatGPT.[104]

LLMs such as ChatGPT gather massive amounts of data from the Internet, including personal information. The Canadian Government has taken a proactive approach towards regulating the use of Generative AI by releasing a draft of a code of practice, which is now open for public comment. The code will be enacted into law as part of the country's Artificial Intelligence and Data Act.[105]

Meanwhile, the G7 has launched the Hiroshima AI Process to coordinate discussions on the risks associated with generative AI.[106] In July 2023, US President Joe Biden announced voluntary commitments from large AI companies to prioritize safety, security, and trust.[107] On July 13, 2023, China implemented temporary measures to regulate the generative AI industry. The new rules mandate that service providers undergo security assessments and file algorithms for review.[108] Additionally, the Beijing Municipal Health Authority has proposed 41 new rules that strictly prohibit the use of AI in various online healthcare activities, including automatically generating medical prescriptions.[109]

102 Perkins Coie (2023). Use of Generative AI in Litigation Requires Care and Oversight, available at: https://perkinscoie.com/insights/blog/use-generative-ai-litigation-requires-care-and-oversight

103 LLM definition by Tech Target: "A large language model (LLM) is a type of artificial intelligence (AI) algorithm that uses deep learning techniques and massively large data sets to understand, summarize, generate and predict new content. The term generative AI also is closely connected with LLMs, which are, in fact, a type of generative AI that has been specifically architected to help generate text-based content", see: https://www.techtarget.com/whatis/definition/large-language-model-LLM

104 McCallum S. (2023). ChatGPT banned in Italy over privacy concerns, available at: https://www.bbc.com/news/technology-65139406

105 Canadian Guardrails for Generative AI – Code of Practice (2023), available at: https://ised-isde.canada.ca/site/ised/en/consultation-development-canadian-code-practice-generative-artificial-intelligence-systems/canadian-guardrails-generative-ai-code-practice

106 G7 Hiroshima Leaders' Communiqué, May 20, 2023, available at : https://www.consilium.europa.eu/media/64497/g7-2023-hiroshima-leaders-communiqu%C3%A9.pdf

107 Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, available at: https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

108 Reuters (2023). China says generative AI rules to apply only to products for the public, available at: https://www.reuters.com/technology/china-issues-temporary-rules-generative-ai-services-2023-07-13/

109 Beijing to limit use of generative AI in online healthcare activities, including medical diagnosis, amid growing interest in ChatGPT-like services, available at: https://www.scmp.com/tech/policy/article/3231828/beijing-limit-use-generative-ai-online-healthcare-activities-including-medical-diagnosis-amid

In other news, the US Federal Trade Commission (FTC) has launched an investigation into OpenAI over allegations of consumer protection law violations. The FTC's Civil Investigative Demand has raised concerns that ChatGPT, a language model developed by OpenAI, may produce false or disparaging statements about real individuals. The agency has also requested information following a data privacy breach in which private user data was exposed in ChatGPT's results.[110]

The example of ChatGPT

ChatGPT (Generative Pre-trained Transformer) is a chatbot that leverages advanced Natural language processing (NLP) and reinforcement learning to participate in realistic discussions with people. ChatGPT can generate articles, tales, poetry, and even computer code. It can also respond to questions, engage in discussions, and, in certain instances, provide extensive replies to extremely precise questions and inquiries. ChatGPT was released in November 2022 and acquired over one million users within a week.[111]

The Judiciary has not been immune from the controversies related to the use of Generative AI. For example, in January 2023 there has been a controversy in Colombia after a judge revealed that he utilized ChatGPT, to help him determine if an autistic child's insurance should cover all expenses related to their medical treatment.[112] Ten days after this controversial ruling, still in Colombia, a Magistrate issued a court order using ChatGPT to help her decide how to conduct a trial in the metaverse. Moreover, in late March 2023, a judge in Peru and a magistrate in Mexico claimed to have used OpenAI's ChatGPT to motivate a second-instance decision and to illustrate their arguments in a court hearing.[113]

Following the Mata vs. Avianca Airlines., Inc case[114], where an attorney submitted falsified citations and cases created by ChatGPT to a US court, responsible use guidelines have become even more essential. The federal judge Brantley Starr (Northern District of Texas) implemented a new rule that demands a more explicit and precise certification. This certification ensures that any text generated by generative AI will undergo a human accuracy check using authoritative legal sources before it is presented to the Court.[115] His order required the following:

> *"All attorneys and pro se litigants appearing before the Court must, together with their notice of appearance, file on the docket a certificate attesting*

110 Reuters (2023). US FTC opens investigation into OpenAI over misleading statements, available at: https://www.reuters.com/technology/us-ftc-opens-investigation-into-openai-washington-post-2023-07-13/

111 https://chat.openai.com/

112 Gutiérrez J. D. (2023). ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the Judiciary, available at: https://verfassungsblog.de/colombian-chatgpt/

113 Gutiérrez J. D. (2023). Judges and Magistrates in Peru and Mexico Have ChatGPT Fever, available at: https://techpolicy.press/judges-and-magistrates-in-peru-and-mexico-have-chatgpt-fever/

114 Mata v. Avianca, Inc., 1:22-cv-01461, available at: https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/

115 Hunton Andrews Kurth (2023). Will Mandatory Generative AI Use Certifications Become The Norm In Legal Filings?, available at: https://www.huntonak.com/en/insights/will-mandatory-generative-ai-use-cerfifications-become-the-norm-in-legal-filings.html. Also see: https://law.mit.edu/ai

either that no portion of any filing will be drafted by generative artificial intelligence (such as ChatGPT, Harvey.AI, or Google Bard) or that any language drafted by generative artificial intelligence will be checked for accuracy, using print reporters or traditional legal databases, by a human being. These platforms are incredibly powerful and have many uses in the law: form divorces, discovery requests, suggested errors in documents, anticipated questions at oral argument. But legal briefing is not one of them. Here's why. These platforms in their current states are prone to hallucinations and bias. On hallucinations, they make stuff up—even quotes and citations. Another issue is reliability or bias. While attorneys swear an oath to set aside their personal prejudices, biases, and beliefs to faithfully uphold the law and represent their clients, generative artificial intelligence is the product of programming devised by humans who did not have to swear such an oath. As such, these systems hold no allegiance to any client, the rule of law, or the laws and Constitution of the United States (or, as addressed above, the truth). Unbound by any sense of duty, honor, or justice, such programs act according to computer code rather than conviction, based on programming rather than principle. Any party believing a platform has the requisite accuracy and reliability for legal briefing may move for leave and explain why. Accordingly, the Court will strike any filing from a party who fails to file a certificate on the docket attesting that they have read the Court's judge-specific requirements and understand that they will be held responsible under Rule 11 for the contents of any filing that they sign and submit to the Court, regardless of whether generative artificial intelligence drafted any portion of that filing".

Source: https://lawprofessors.typepad.com/appellate_advocacy/2023/10/disclosing-the-use-of-ai.html

These are three main risks of generative AI regarding judiciaries:

- **Purpose/scope creep**. An AI system designed and deployed for purpose "A" should not blindly be used for some alternative function. For example, an NLP tool primarily for translation of court orders should not arbitrarily be used for also aiding case queries or assist judges in decision making without disclosing of its usage for such additional purposes. In some instances, the additional purposes may be valid, in others not. Even where additional functions can be deemed legal and valid, it may be necessary to train the base algorithm on additional relevant data to ensure accuracy and reliability. Basically, a blind expansion of purpose creep generally exacerbates the potential risks of a general-purpose AI system and should be deterred or at least regulated.

- **Hallucinations and mis/disinformation.** It's important to keep in mind that generative AI models are trained on extensive amounts of data, resulting in highly realistic and relevant responses. However, it's worth noting that tools utilizing such models may produce outputs that are

plausible but not entirely accurate due to the nature of their design, which aims to generate output that closely resembles but may not be identical to the source information. General purpose AI, especially LLMs have been increasingly demonstrating a potential to "hallucinate" - i.e., give out inaccurate outputs in a compelling human-like manner thus, making them credible and increasing the risk of their acceptance as accurate (a form of automation bias). This is particularly dangerous in the judicial system - we have had different instances in the last few months of judges relying on ChatGPT to give inputs on existing jurisprudence regarding legal question. This was reported in Colombia in an insurance case, and even in India (Punjab & Haryana High Court judge). Hallucinated output can prove extremely problematic, especially for adjudication.

- **Intellectual property concerns.** LLMs again need to be considered given the concerns around traditional IP rights of original work creators.

Activity: Training participants read the text below on copyright implications of using Generative AI and discuss if the doctrines of "fair use" or "permissible copyright exceptions" could be applied in the context of Generative AI?

With the rise of generative AI, lawsuits seem to be becoming a daily occurrence. In November, 2022, Microsoft, GitHub, and OpenAI faced a class action lawsuit alleging that the Copilot system owned by GitHub, which was trained on billions of lines of public code, violates copyright law by regurgitating licensed code snippets without attribution.[116] In return, the companies argued before a federal court in San Francisco that the current lawsuit regarding their use of open-source code to train their AI systems is not sustainable. The companies asserted that the complaint lacks specificity in its allegations. Additionally, they argued that GitHub's Copilot system, which provides code suggestions to programmers, utilizes the source code in a manner consistent with fair use principles.[117]

There is also a legal case against Midjourney and Stability AI, the companies responsible for widely-used AI art tools. The case claims that these companies violated the rights of millions of artists by using web-scraped images to train their tools.[118]

Moreover, Getty Images filed a lawsuit against Stability AI for allegedly utilizing millions of images from their site without authorization to train Stable Diffusion, an AI capable of generating art.[119]

116 Vincent J. (2022). The lawsuit that could rewrite the rules of AI copyright, available at: https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

117 IT world Canada (2023). Microsoft, GitHub, and OpenAI ask court to dismiss AI copyright lawsuit, available at: https://www.itworldcanada.com/post/microsoft-github-and-openai-ask-court-to-dismiss-ai-copyright-lawsuit

118 Vincent J. (2023). AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, available at: https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

119 Brittain B. (2023). Getty Images lawsuit says Stability AI misused photos to train AI, available at: https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

The main concern with generative AI is its inclination to replicate images, text, and other types of content, including those that are copyrighted, from its training data. This issue was highlighted in a recent incident where an AI tool utilized by CNET for writing explanatory articles was discovered to have plagiarized human-written articles, which were likely part of its training dataset.[120] Additionally, a December academic study revealed that AI models capable of generating images, such as DALL-E 2 and Stable Diffusion, can replicate certain elements of images from their training data. [121]

Certain platforms that host images have prohibited the use of AI-generated content due to potential legal repercussions. Legal professionals have also warned that using generative AI tools may expose companies to risk if they inadvertently integrate copyrighted content produced by these tools into their products for sale.

It has been argued by companies such as Stability AI and OpenAI, the creators of ChatGPT, that they are protected by the "fair use" doctrine even if their systems were trained using licensed content. This legal principle, which is recognized in the United States, allows for the limited use of copyrighted material without obtaining permission from the owner of the rights. Fair use advocates often cite the example of Authors Guild v. Google, where the U.S. Court of Appeals for the Second Circuit in New York determined that Google's manual scanning of millions of copyrighted books to develop its book search platform was a fair use, even without a license. However, the concept of fair use is frequently debated and modified, and it remains largely untested in the realm of generative AI.[122]

Whether the works produced by AI can be protected by the 'fair use' defense depends on whether they are deemed transformative. This means that the works must use copyrighted materials in a way that significantly differs from the originals. Past legal cases, such as the Google v. Oracle decision from the US Supreme Court in 2021, indicate that creating new works from gathered data can be transformative. The court found that Google's use of parts of Java SE code to develop its Android operating system was considered fair use.[123]

Source: Tech Crunch (2023). The current legal cases against generative AI are just the beginning, available at: https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/

120 Futurism (2023). CNET's AI Journalist Appears to Have Committed Extensive Plagiarism, available at: https://futurism.com/cnet-ai-plagiarism

121 Somepalli G., Singla V., Goldblum M., Geiping J., Goldstein J. (2022). Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, University of Maryland, available at: https://arxiv.org/pdf/2212.03860.pdf

122 Authors Guild v. Google, Inc., No. 13-4829 (2d Cir. 2015), available at: https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html

123 Setty R. (2023). First AI Art Generator Lawsuits Threaten Future of Emerging Tech, available at: https://news.bloomberglaw.com/ip-law/first-ai-art-generator-lawsuits-threaten-future-of-emerging-tech

The use of AI techniques can reduce the requirement for human translation. These tools can rapidly identify documents containing foreign language text and provide a list of the languages they contain, enabling more thorough planning. Several AI technologies can also translate text from one language to another.

Natural Language Processing (NLP)

NLP is an ML technique that analyses vast amounts of human text or speech data (transcribed or acoustic) for specific properties, such as meaning, content, intention, attitude, and context[124].

Language analysis has been used in the legal domain and criminology for a long time. For instance, text classification has been used in forensic linguistics. While in the past the analysis was done manually, today ML methods are used to identify gender, age, personality traits, and even the identity of an author, or for live transcription.[125] For instance, NLP can support judicial operators in identifying and linking references to the same person or organization throughout a set of legal contracts. It can also be used in analysing a collection of court cases to identify recurring legal topics or issues, or in extracting the names of parties involved, dates, and locations mentioned in a court opinion. Moreover, NLP systems can be used to automatically redact sensitive information from court documents, such as Social Security numbers and personal addresses, to protect individuals' privacy.

It should be noted that NLP models are still error prone, and errors in translation can have serious consequences for fundamental rights of individuals when these models are deployed in judicial operations.

124 Firth-Butterfield K., Silverman K. (2022). Artificial Intelligence and the Courts: Materials for Judges. Artificial Intelligence – Foundational Issues and Glossary, American Association for the Advancement of Science, available at https://doi.org/10.1126/aaas.adf0782

125 Medvedeva M., Vols M., Wieling M. (2020). Using machine learning to predict decisions of the European Court of Human Rights, Artif Intell Law, 28, 237–266, available at: https://link.springer.com/article/10.1007/s10506-019-09255-y

Case Study

India's SUVAS

The Vidhik Anuvaad Software (SUVAS), an AI program that translates decisions and orders into nine different local languages, was introduced by the Supreme Court in November 2019. SUVAS aimed to make it easier for people who do not speak English to obtain judgements and orders and to help them to gain a better understanding of court proceedings.

Source: Press Trust of India, Software developed to translate SC judgments in 9 vernacular languages: Law Minister RS Prasad, available at: https://www.business-standard.com/article/pti-stories/software-developed-to-translate-sc-judgments-in-9-vernacular-languages-law-minister-rs-prasad-119121200851_1.html.

In February 2023, Technology Enabled Resolution (TERES), a tech startup based in Bangalore, India, started using AI to start live transcription of Supreme Court hearings.

Source: Mint (2023). Bangalore techies bring AI to Supreme Court for the first time, available at: https://www.livemint.com/news/india/supreme-court-uses-ai-based-transcript-for-the-first-time-here-s-how-it-works-11677403522929.html.

IIndia has been successful in creating its own NER and Rhetorical Role Models trained on Indian Legal text. The NER model specifically is at 91% accuracy.

Source: https://github.com/OpenNyAI/Opennyai

Digital file and case management

AI could also facilitate digital file management, which in turn, would make judicial operators more effective by enabling them to focus on more substantive matters.

Intelligent Trial 1.0, a smart court management AI in China

For instance, the Hebei High Court in China has developed an Intelligent Trial 1.0, a smart court management AI. It automatically scans and digitizes filings; classifies documents into electronic files; matches parties to existing case parties; identifies relevant laws, cases, and legal documents to be considered; generates all necessary court procedural documents such as notices and seals; and distributes cases to judges so that they can be put on the right track. The technology coordinates numerous AI tasks into a workstream that can minimize the burdens of court personnel and judges.

Tool for the anonymization of legal documents, Argentina

To speed up the judicial process and lower the margin of error, Cambá Cooperative, a software bank labour cooperative, has created a scalable AI system to anonymize legal papers in Spanish. The AI system aims to anonymize personal data of public documents, reduce time and errors, and safeguard the right to privacy. The Criminal court no. 10 in Buenos Aires, Argentina implemented this AI tool in their rulings.[126]

In depth: AI as evidence in judicial proceedings

The complex nature of ML algorithms and their opaque nature pose challenges to using AI systems as evidence in legal proceedings. Courts must establish a reliable method to verify the accuracy of AI outputs, which may involve expert testimony or technical means like watermarks embedded in images. Deciding who is qualified to testify on the accuracy of AI applications is also a crucial issue, with options ranging from software engineers and design engineers to data engineers and company CEOs.[127]

Judges face difficulties in determining the accuracy of AI-powered diagnostic tools. While medical diagnostic AI can be compared to physician diagnoses, It is unclear how algorithms designed to predict future behaviour, such as criminal assessment tools, can be scientifically or evidentially evaluated. It can be challenging to determine causation with predictive algorithms in the criminal context, as they also consider social factors that may influence behaviour. Assessing accuracy, error rates, and conducting testing and peer review are crucial but difficult tasks in this field. Once a person has been incarcerated or sentenced, it becomes difficult to predict how their future behavior may have been influenced by their imprisonment. The effects of imprisonment, including the support of loved ones on the outside, can have a significant impact on a person's future behaviour, making it extremely difficult to accurately gauge the accuracy of ML's prediction.

Litigating parties will also seek to challenge the relevance and accuracy of the ML system by seeking access to the underlying algorithm, the data on which it was trained, validated, and tested, as well as what occurs and is weighted inside any machine-learning black box. Thus, courts could face layered adjudicative challenges each time AI generated evidence is offered. Where AI outputs are admitted, opponents will seek to cross examine the software engineers responsible for its design. Moreover, because each AI application is different, i.e., it will:

- Have different output purposes;
- Rely on different algorithms;
- Use different machine learning methodologies;
- Train, test, and validate using different data.

These issues are generally not subject to resolution through the application of case law precedent in the same way, for example, that DNA analysis is now generally accepted in court. One should expect the adjudication of each application and in each context for which the application is offered as evidence.

126 See: https://www.empatia.la/en/proyecto/ia2/; see also: Selvood I., Uribe P. (2022). Open Justice is Moving Forward in the Americas, available at: https://www.opengovpartnership.org/stories/open-justice-is-moving-forward-in-the-americas/

127 Baker J. E., Hobart L. N., Mittelstead M. G. (2021). AI for Judges. A Framework. Center for Security and Emerging Technology, available at: https://www.armfor.uscourts.gov/ConfHandout/2022ConfHandout/Baker2021DecCenterForSecurityAndEmergingTechnology1.pdf

The rapid advancements in AI and NLP technologies present new possibilities for modernising the justice sector in Africa. For instance, companies like Juta[128] in South Africa are leveraging these innovations to develop cutting-edge solutions that aid law firms and other legal organisations in conducting comprehensive legal research and unearthing valuable resources for their cases[129]. By capitalising on Juta's vast repository of legal documents and utilising advanced analytical techniques, African judicial systems can enhance their efficiency and effectiveness.

Case Data

One potential area where AI technology could be incorporated into African judicial systems is through digitalisation of court case data. By capturing detailed information on various aspects of the legal process - including judgments, rulings, decisions, case backgrounds, parties involved, etc - it would enable deep learning algorithms to identify patterns and insights from this data. Properly organising and storing the gathered case data into large databases will help set a foundation enabling Africans to leverage its value for a multitude of applications and actionable features. The recording system's accuracy must also be auditable with physical proof given priority over digital records. This is not an easy task and will require great coordination in the judicial system. Analyses as straightforward as tracking the different courts progress vs each judge's predecessors in previous years could determine which judge to assign certain types of trials based on productivity under average target periods by high success rate correlation studies. Another sphere would be examining legal statements from within a case which is open source or government generated datasets.

Handling discovery and information retrieval

To improve the efficiency of the discovery phase in legal proceedings and facilitate more effective sharing of relevant documentation among stakeholders, the implementation of digital archives is crucial[130]. By establishing an online platform for storing essential files and evidence, judicial systems can take advantage of state-of-the-art search mechanisms to locate critical pieces of information quickly and accurately. Such an approach not only streamlines information management but also enables attorneys to build stronger arguments supported by reliable facts derived from accessible, interconnected sources.

128 Juta and Company is a leading provider of quality legal, regulatory, business and academic content across Africa; see: https://juta.co.za

129 Jutastat Evolve is a cognitive analytical research solution for fast, accurate discovery, data insights and analytics; see https://jutastatevolve.co.za/

130 Kufakwababa C. Z. (2021). Artificial intelligence tools in legal work automation: The use and perception of tools for document discovery and privilege classification processes in Southern African legal firms, Doctoral dissertation, Stellenbosch: Stellenbosch University.

Using multimodal court proceedings

Modernising courtroom environments through multipurpose media gathering could greatly benefit judicial operations throughout Africa. Integrating a range of sensory inputs, including audio and video recordings, offers several advantages. Technological advances in computer vision and machine listening can substantially enhance how transcribers convert spoken words to text, diminishing human error while boosting speed and precision. These digital transcripts become tools for post-trial investigative analysis, and when combined with predictive modelling capabilities, pave the way for more sophisticated decision support during active hearings. Furthermore, indexing multimedia assets for easy access facilitates both Judiciary reference and public scrutiny, contributing to increased trustworthiness within the legal framework. Decision makers might consider implementing pilot projects using smart recordkeeping methods and observing promising outcomes before wider adoption of open-ended multimedia formats. Then systemic changes can be catered to specific national priorities.

Improving language tools for the Judiciary

Using advanced natural language processing technologies like Machine Translation[131] and Document Classification provides an excellent opportunity for African judiciaries to address linguistic barriers. Lack of local language support hinders public engagement and dissemination of vital information regarding judicial proceedings. Adopting modern AI solutions for translations ensures equitable access to legal resources across diverse populations with varying native tongues. Meanwhile, classifying local language content empowers the justice system to accept and analyse multi-cultural submissions, narrowing geographical divides between interpreters, litigators, and court personnel. For instance, academic reports published by professional associations stress that eliminating linguistic discrimination and promoting parity within the legal domain could alleviate similar issues related to jurisprudence around Africa[132]. Given increased interest directed towards developing regional lexicons and inferential techniques, more nations can capitalise upon tailored constituent presentations. Subsequently, governments would demonstrate tangible commitment towards constructively integrating remote areas.

131 Adelani D., Alabi J., Fan A., Kreutzer J., Shen X., Reid M., Ruiter D., Klakow D., Nabende P., Chang E., Gwadabe T. (2022). A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation, In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3053–3070.

132 Docrat Z., (2022). A Review of Linguistic Qualifications and Training for Legal Professionals and Judicial Officers: A Call for Linguistic Equality in South Africa's Legal Profession, International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique, 35(5), 1711–1731

Open Legal Repositories

Open repositories containing comprehensive collections of continental judicial decisions serve as valuable resources for legal practitioners and academics alike. Aspects like ease of navigation amplify the importance of these databases in promoting informed deliberations. While some African nations have made significant strides in digitising their high-court rulings, lower courts remain comparatively underrepresented. Despite having dedicated legal tech organizations, such imbalanced distribution warrants attention. Therefore, enhancing information technology infrastructure for judicial institutions becomes necessary to ensure uniform coverage of all courts, fostering balanced accessibility and equal opportunities for advancement via legal data insights.

Connecting with local AI

African academic institutions and private research facilities focusing on Artificial Intelligence (AI) should be sought after by the Judiciary to bolster joint collaborations that maximise benefits stemming from these partnerships. Encouragement of these interactions helps navigate complex international regulatory frameworks via shared expertise and experience. Additionally, involvement with prolific homegrown AI community initiatives like the Deep Learning Indaba, Data Science Africa, Masakhane Research Organisation, Data Science Network, which comprises numerous researchers spread across various nations, could greatly enhance judicial systems' connection to innovative minds within the region. Thus, embracing continent-wide collaboration possesses transformative potential spanning technical proficiency within courts and general societal inclusion.

2. Case studies on AI deployment in the Judiciary

This section gives a general overview of select cases of AI deployment in the Judiciary in Brazil, Singapore, Argentina, Colombia, India, UK and US. It must be noted that this does not serve as an endorsement of these use cases of AI in select national judiciaries, and that judicial operators need to be aware of all risks (bias, black boxes, cybersecurity, and encroachment of human rights) that might occur with the use of AI systems in judicial operations.

VICTOR, Brazil

The Brazilian Supreme Court (STF) uses the VICTOR AI system, which was developed in collaboration with the University of Brasilia (UnB). The AI technology analyses the enormous volume of appeals brought to the High Court and automates the examination process by identifying cases with repercussão geral (general repercussion), a requirement for the processing of an appeal before the STF.

Only in 2018, more than fifty thousand appeals were filed with this Court, which has the potential to decide around one hundred and twenty thousand cases annually. The first stage in analyzing all appeals that reach the STF is determining whether they have general repercussions. Before VICTOR, this analysis was performed by court officials based on the binding precedents of the Justices, and it took around forty minutes per case.

Regarding its software design, VICTOR incorporates various cutting-edge technologies and a vast database of court documents. The dataset used to train VICTOR contains more than 100,000 lawsuits and nearly three million case dockets extracted over a two-year period (2017-2019).

Its initial problem was to deal with the reality that court documents from all Brazilian courts (State, Federal, Labour, Military, Electoral Justice) arrive at the STF in varied formats, such as unstructured PDF volumes containing unindexed documents.[133]

Singapore's Intelligent Court Transcription System

The Intelligent Court Transcription System (iCTS) has been implemented in Singapore courts in partnership with A*STARs Institute for Infocomm Research. The iCTS has the potential to increase court efficiency by transcribing court hearings in real-time, removing the need to hire a human transcriber and allowing judges and parties to review oral testimonies

133 Salomao L. F., Braga R. (2020). The role of the Judiciary in the realization of the UN 2030 Agenda, available at: https://www.conjur.com.br/2021-jul-09/salomao-braga-judiciario-agenda-2030-onu/. See also: https://portal.fgv.br/en/news/artificial-intelligence-Judiciary-and-its-role-implementing-un-agenda-2030; https://sifocc.org/app/uploads/2020/06/Victor-Beauty-or-the-Beast.pdf

in court immediately. It does this by using neural networks trained with language models and domain-specific terms (such as legal terminology).[134]

It has to be noted that speech recognition systems have a "reputation" of not performing well when exposed to certain accents, which ends up being discriminatory under certain circumstances. Judicial operators need to be aware of these shortcomings.

Prometea, Argentina

The Prometea system uses AI approaches to generate court opinions automatically. In 2017, the Prosecutor's Office in the Autonomous City of Buenos Aires, Argentina, began developing Prometea. The tool has enabled the Prosecutor's Office to significantly improve the efficiency of its processes: a reduction from 90 minutes to one minute (99%) for the resolution of a tender process,and from 167 days to 38 days (77%) for trial preparation.[135]

Prometea is distinguished by three primary characteristics:

- It offers an intuitive and user-friendly interface that enables natural language recognition and "talking" to the machine. On a single screen, the user has access to all their work-related resources.
- It operates as a multifunctional expert system with the ability to automate document processing and provide intelligent support.
- It employs supervised ML and clustering approaches, based on hand labelling and training on machine-generated datasets.[136]

The functionalities of Prometea can be divided into four categories:

- Intelligent Assistance: Prometea aids decision makers and users in achieving a result using its voice or a chatbot. The system automates tasks associated with the deadline control of filed judicial appeals; analyses the relevant paperwork accompanying the file and with a query based system with just five questions, judges can develop a legal opinion to decide on an appeal.
- Automation: the concept of automation has different subtleties based on numerous circumstances. There are mostly two large groups:
    - Complete Automation: the algorithms automatically associate data and information with documents. The document is generated without interaction from a person.

134 Lee J. (2020). Legal Tech-ing Our Way to Justice, available at: https://lawtech.asia/legal-tech-ing-our-way-to-justice/. See also: https://www.a-star.edu.sg/docs/librariesprovider10/default-document-library/fw-new-infosheets/smart-nation-digital-economy/intelligent-court-transcription-system.pdf?sfvrsn=72a5a971_3

135 UNESCO Chair on Knowledge Societies and Digital Government (2020). PROMETEA: Transforming the administration of justice with artificial intelligence tools, available at: https://unescochair.cs.uns.edu.ar/en/2020/06/prometea-transforming-the-administration-of-justice-with-artificial-intelligence-tools/. See also: Corvalan J. G., Le Fevre Cervini E. M. (2020). Prometea experience. Using AI to optimize public institutions, available at: https://ceridap.eu/prometea-experience-using-ai-to-optimize-public-institutions; https://www.ibanet.org/article/14AF564F-080C-4CA2-8DDB-7FA909E5C1F4

136 Corvalan J. G., Le Fevre Cervini E. M. (2020). Prometea experience. Using AI to optimize public institutions, available at: https://ceridap.eu/prometea-experience-using-ai-to-optimize-public-institutions

- Automation with Reduced Human Intervention: In many instances, human interaction with an automated system is required to complete or enhance the generation of a document.

- Intelligent Classification and Detection: The detection is based on the reading and analysis of a massive volume of information, in which Prometea may identify documents based on different combinations of criteria, regardless of the documents' linguistic diversity. Then, the system segments data based on shared patterns (keywords) throughout the documents.
- Prediction: It is the most complex function offered by Prometea. A prediction will be made based on past responses. When Prometea finds a match between the present document and a previous one, it keeps note of the answer provided in previous situations and suggests the same remedy because the conditions are similar. This work is derived from the reading and recognition of precedent web-accessible judicial decision patterns from prior instances. Once Prometea identifies the solution, it allows the user to complete the legal opinion based on a few questions and then displays an online-editable preview of the final document. The first draft of the document is generated automatically by the AI system.[137]

Given the ongoing concerns regarding the justification of Prometea's decisions and their implications for due process, civil society has urged for sustained oversight of the program's execution. Other issues to be cautious about are the level of accountability of the relevant players (developers and judges), and potential biases in training data and design.[138]

PretorIA, Colombia

At the start of 2019, the Colombian Constitutional Court announced a pilot project of implementing Prometea to resolve the inefficiency and backlogs. Every day, the court receives more than 2,000 writs of protection from all the courts across the nation. Only nine judges and fewer than 200 staff members work for the Constitutional Court. However, academics and members of civil society raised numerous concerns about Prometea's potential effects, as well as its operation and the decision-making process that were considered opaque. Prometea turned out to be a pilot that has been put on hold. The biggest challenge was related to privacy and data protection related to the sharing of sensitive information with third parties, such as software

137 Ibid.
138 OECD, AI use cases in LAC governments, available at: https://www.oecd-ilibrary.org/sites/08955f48-en/index.html?itemId=/content/component/08955f48-en

developers. It is crucial that victims' identity and their personal information or data be protected in cases where minors are involved, or in cases where sexual offences are implicated, among other circumstances. Access to this information or data by anybody other than the court and those parties involved in the processing of cases constituted a violation of confidentiality. Given the system's weakness in this regard, it was especially concerning that a potential leak of personal information to the media or other interested parties could happen, with potentially disastrous results for the protection of privacy for persons engaged in the cases processed by the AI system.[139]

Following multiple debates, the Constitutional Court changed the project by implementing clearer and more transparent technology. PretorIA, released in the middle of 2020, uses topic modelling technology rather than neural networks because of this. The new version can be completely explained, interpreted, and tracked.[140]

SUPACE, India

The Indian Judiciary has a large number of pending cases. According to data from the National Judicial Data Grid, around 38 million cases are outstanding in various district and taluka courts in India, and over one hundred thousand cases have been pending for more than three decades.[141]

The Supreme Court of India has implemented an AI system, Supreme Court Portal for Assistance in Courts Efficiency (SUPACE) that will aid in the administration and delivery of justice through cataloguing a large number of earlier judicial decisions for better processing of case material, whether for comprehending the factual matrix of specific instances or conducting dynamic research into precedents. SUPACE will not be used in decision-making. The role of AI will be confined to data collection and analysis.[142]

The SUPACE AI tool is being deployed on an experimental basis with judges handling criminal cases at the Bombay and Delhi High Courts.

The Supreme Court of India is exploring the use of a mobile application that will translate the court's decisions into nine languages. In addition, India is using AI to resolve minor charges such as traffic violations.[143]

139 Guitierrez O. L. C., Castañeda J. D., Saavedra Rionda V. P. (2019). Enthusiasm and complexity: Learning from the "Prometea" pilot in Colombia's judicial system, available at: https://giswatch.org/node/6166
140 Ibid.
141 Shanthi S. (2021). Behind SUPACE: The AI Portal Of The Supreme Court of India, available at: https://analyticsindiamag.com/behind-supace-the-ai-portal-of-the-supreme-court-of-india/
142 Ibid.
143 Ibid.

HART (Harm Assessment Risk Tool), United Kingdom

The Harm Assessment Risk Tool ('HART') is used by Durham Constabulary in the United Kingdom. Using more than thirty characteristics that describe a person's criminal history and socioeconomic background, HART uses an ML algorithm to determine a suspect's probability of reoffending. The local police use the risk assessments completed by HART to decide whether to charge a person or divert them into a rehabilitation programme. HART does not decide whether a person is guilty or innocent, but its evaluation may start a series of actions that lead to a person being deprived of their freedom or being found guilty of a crime. Charges should undoubtedly be determined by the merits of each individual case, and it is difficult to see how judgments about participation in rehabilitation programmes could be decided in any other way than by carefully analyzing each person's unique situation. There should always be a human in the loop overseeing the output of an automated decision-making system that makes high-impact and fact-sensitive decisions.[144]

HART is prone to over-criminalize since it is intentionally meant to underestimate who is qualified for enrolment into the rehabilitation programme. This method runs counter to the idea that every ambiguity in a criminal case should be resolved in the defendant's favour ("in dubio reo"). Contrary to what HART does, a human rights compliance approach to criminal justice decision-making would need to favour the defendant.[145]

144 Oswald M., Grace J., Urwin S., Barnes G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, Information & Communications Technology Law, 27 (2), 223–250, available at: https://doi.org/10.1080/13600834.2018.1458455

145 Ibid.

3. Activities

The following group activities are intended to encourage the training participants to discuss various implications related to the use of AI in the Judiciary

Activity 1

Please discuss the following questions with other training participants:

- Who should be accountable for automated decisions and how should responsibility be allocated within the chain of actors when AI facilitates the final decision?
- What is a fair trial if ADM has facilitated the decisions?
- Is accused denied due process of law when AI systems are deployed at some stage of the criminal procedure?

Activity 2

Please go to the following link: https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment#illustration. The illustration "demonstrates how risk scores are calculated in risk assessment. For the sake of illustration, this hypothetical example only covers five domains of predictors, including demographics, criminal history, education/employment, family/social support, and antisocial cognition, and only one indicator for each domain. Values on each indicator have been assign ed scores ranging from 0 to 2; the higher the score, the more likely one is to reoffend (e.g., because younger persons are more likely to reoffend than older persons, values on the "age at sentencing" indicator decrease as age increases)."[146]

Enter certain characteristics to understand better how the risk assessment tool works. Discuss with other training participants what its advantages and disadvantages are.

Activity 3

Training participants read the hypothetical scenario: "Navigating the Risks: Judges Using Generative AI" and discuss the key challenges in deploying Generative AI by courts.

Scenario Description:

In a future where generative AI has made significant advancements, judges have started to experiment with its use in the courtroom. However, they

146 https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment

soon encounter several challenges and risks associated with its adoption. This scenario highlights the potential risks and pitfalls of using generative AI in a judicial context.

Scenario Elements:

1. **Automated Legal Document Generation:**
   - Judges begin using generative AI to automate the drafting of legal documents, such as judgments and opinions.
   - The AI system, while efficient, sometimes generates biased or inaccurate legal arguments and conclusions.
2. **Overreliance on AI Assistance:**
   - Judges increasingly rely on AI-generated legal analysis, gradually reducing their own critical thinking and decision-making skills.
   - There is a growing concern that judges may become passive users of AI, diminishing their role in interpreting and applying the law.
3. **Ethical and Legal Bias:**
   - The AI models used by judges inherit biases present in their training data. This leads to decisions that disproportionately favor certain groups or perpetuate existing biases in the legal system.
   - Legal scholars and activists raise concerns about fairness and discrimination.
4. **Transparency and Accountability:**
   - Generative AI models can be complex and difficult to interpret. Judges face challenges in explaining AI-generated decisions to litigants, attorneys, and the public.
   - Questions arise about the accountability of AI-generated decisions, particularly in cases where they result in adverse consequences.
5. **Data Privacy and Security:**
   - The use of generative AI in court proceedings involves handling vast amounts of sensitive legal data. Concerns emerge about data breaches and the security of confidential information.
   - Courts must invest heavily in cybersecurity to protect against potential threats.
6. **Public Trust and Perception:**
   - As generative AI becomes more integral to the legal process, public trust in the justice system is eroded.

- Citizens and litigants express skepticism about the fairness and impartiality of AI-assisted decisions.

7. **Legal Challenges and Precedents:**
   - Legal challenges arise over the admissibility of AI-generated evidence and whether AI can be considered a reliable source of legal analysis.
   - Courts are faced with the task of establishing legal precedents to govern the use of AI in their decisions.

Scenario Outcome:

As judges grapple with the risks and challenges associated with the use of generative AI in the courtroom, they must carefully balance the potential benefits of efficiency and accuracy with the need to preserve transparency, fairness, and human judgment in the legal system. The scenario underscores the importance of comprehensive guidelines, oversight mechanisms, and ongoing training to mitigate these risks and ensure that AI enhances, rather than undermines, the principles of justice.

1. AAAS, Artificial Intelligence and the Courts: Materials for Judges, available at: https://www.aaas.org/ai2/projects/law/judicialpapers
2. Abu Elyounes D. (2019). Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness, Journal of Law, Technology and Policy, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3478296
3. Ada Lovelace Institute, AI Now Institute and Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/
4. Bhuiyan J. (2021). LAPD ended predictive policing programs amid public outcry. A new effort shares many of their flaws, available at: https://www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform
5. Brittain B. (2023). Getty Images lawsuit says Stability AI misused photos to train AI, available at: https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/Council of Europe, available at: https://www.coe.int/en/web/cepej
6. Council of Europe (2021). CEPEJ Action plan 2022 – 2025: "Digitalisation for a better justice", available at: https://rm.coe.int/cepej-2021-12-en-cepej-action-plan-2022-2025-digitalisation-justice/1680a4cf2c
7. Futurism (2023). CNET's AI Journalist Appears to Have Committed Extensive Plagiarism, available at: https://futurism.com/cnet-ai-plagiarism
8. Hind M. (2019). Explaining Explainable AI by Michael Hind, The ACM Magazine for Students, 25(3), available at: https://doi.org/10.1145/3313096
9. Jauhar A., Misra M., Sengupta A., Chakrabarti P. P., Ghosh S., Ghosh K., (2021). Responsible Artificial Intelligence for the Indian Justice System, available at: https://vidhilegalpolicy.in/wp-content/uploads/2021/04/Responsible-AI-in-the-Indian-Justice-System-A-Strategy-Paper.pdf
10. IT world Canada (2023). Microsoft, GitHub, and OpenAI ask court to dismiss AI copyright lawsuit, available at: https://www.itworldcanada.com/post/microsoft-github-and-openai-ask-court-to-dismiss-ai-copyright-lawsuit
11. Somepalli G., Singla V., Goldblum M., Geiping J., Goldstein J. (2022). Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, University of Maryland, available at: https://arxiv.org/pdf/2212.03860.pdf
12. The Surveillance and Policing of Looted Land (2021). Automating banishment, available at: https://automatingbanishment.org/section/2-architecture-of-data-driven-policing/
13. UNESCO MOOC on AI and the Rule of Law, available at: https://www.unesco.org/en/articles/unesco-global-mooc-ai-and-rule-law-engaged-thousands-judicial-operators
14. UNESCO (2021). Global Toolkit for Judicial Actors: International legal standards on freedom of expression, access to information and safety of journalists, available at: https://unesdoc.unesco.org/ark:/48223/pf0000378755

15. Vincent J. (2022). The lawsuit that could rewrite the rules of AI copyright, available at: https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

16. Vincent J. (2023). AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, available at: https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

17. Završnik A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings, European Journal of Criminology, 18, 623–642, available at: https://doi.org/10.1177/1477370819876762

Module 3

Legal and Ethical Challenges of AI

Module three discusses the legal and ethical risks associated with AI systems, and the challenges of algorithmic transparency and accountability in the Judiciary. It then proceeds with an overview of the most salient legal issues related to biometric identification and facial recognition technology. The module also elaborates on the key challenges related to AI and ethics based on the UNESCO 2021 Recommendation on the Ethics of Artificial Intelligence.

What will you learn?

After completion of this module, the participants will be able to:

- Understand and explain key challenges related to algorithmic transparency and accountability in the Judiciary, and relevant case law;
- Understand the most salient legal issues related to biometric identification, facial recognition technology and deep fakes;
- Have a firm grasp of the key challenges related to AI and ethics based on the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021).

1. What is AI Ethics?

The UNESCO Recommendation on the Ethics of AI, approaches AI ethics as a systematic normative reflection, based on a holistic, comprehensive, multicultural and evolving framework of interdependent values, principles and actions that can guide societies in dealing responsibly with the known and unknown impacts of AI technologies on human beings, societies and the environment and ecosystems, and offers them a basis to accept or reject AI technologies.

UNESCO considers ethics as a dynamic basis for the normative evaluation and guidance of AI technologies, referring to human dignity, well-being and the prevention of harm as a compass and as rooted in the ethics of science and technology.

In practice, the ethical AI involves considering the ethical implications of AI systems and ensuring that their design and implementation align with broader societal values and norms.

Thought experiment

Let's try a thought experiment: You are at the tram stop and suddenly notice a trolley speeding towards five individuals who are unaware of its approach. You also see a second track that has only one person on it. What would you do? Would you choose to divert the trolley to the second track to save the five individuals at the cost of one life?

For many years, the trolley problem has been a renowned ethical dilemma tackled in philosophy courses. However, the emergence of experimental self-driving cars has brought this theoretical problem to reality. As a result, we are now faced with the challenge of determining the appropriate programming for AI systems in critical life-or-death situations.

Source: Utrecht University, Unboxing the black box of AI, available at: https://www.uu.nl/en/organisation/in-depth/unboxing-the-black-box-of-ai

Many self-regulatory initiatives have focused on the ethical risks posed by AI. Governments, international organizations, the private sector, civil society organization, have all produced non-binding ethical rules and principles to guide the development and use of AI. This chapter gives an overview of key AI ethical frameworks, focusing on the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021). It is important to note that the UNESCO Recommendation as well as other ethics frameworks on AI do not have the binding effects of law.

Table 4. below gives an overview of key principles of AI ethics.

Table 4. Key AI Ethics Principles

Principles	Explanation
Fairness and Bias	AI systems should be designed to ensure fairness and avoid biases that may lead to discriminatory outcomes. It is crucial to address biases in training data, algorithms, and decision-making processes to prevent unjust treatment or marginalization of certain individuals or groups.
Transparency and Explainability	AI systems should be transparent, providing users with an understanding of how they work and how decisions are made. Explainability is important to ensure accountability, enable auditing, and build trust in AI technologies (see, NIST guidance, p. 43).
Privacy and Data Protection	AI systems often rely on vast amounts of data, including personal and sensitive information. Respecting privacy rights and adhering to data protection regulations are essential in AI development and deployment. Minimizing data collection, ensuring informed consent, and safeguarding data from unauthorized access are key considerations. Respecting, protecting, and promoting privacy is crucial for safeguarding human dignity, autonomy, and agency throughout the entire life cycle of AI systems.[147]
Accountability and Responsibility	Clear lines of accountability should be established for the outputs of AI systems, including identifying who is responsible for the actions and decisions made by AI technologies. Ensuring that there are mechanisms for redressing the potential negative impacts of AI systems is crucial.
Safety and Robustness	AI systems should be designed with safety in mind to prevent unintended harm. Measures should be taken to ensure that AI technologies are robust, reliable, and able to handle unforeseen circumstances and adversarial attacks.
Human Autonomy and Oversight	AI should be developed and used to enhance human autonomy and decision-making, rather than replacing or unduly influencing human judgment. Maintaining human oversight and control over AI systems is important to preserve human agency. It is crucial to make sure that ethical and legal responsibility can be assigned to physical persons or existing legal entities at every stage of the AI system›s life cycle. This includes cases where remedies are needed. Human oversight means more than just individual supervision; it also involves inclusive public monitoring as needed.[148]
Social, Environmental and Economic Impacts	AI technologies can have profound social and economic impacts. Ethical considerations include ensuring equitable access to AI benefits, minimizing job displacement, and addressing broader societal implications such as wealth inequality and the digital divide.

147 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137
148 Ibid.

Principles	Explanation
Inclusiveness and diversity	It is crucial to prioritize respect, protection, and promotion of diversity and inclusiveness when developing AI systems, in accordance with international law and human rights. This can be achieved by encouraging the active participation of all individuals and groups, irrespective of their race, colour, descent, gender, age, language, religion, political views, national or ethnic origin, social or economic background, disability, or any other factors.[149]
Collaboration and Multidisciplinary Approaches	Addressing AI ethics requires collaboration among various stakeholders, including researchers, policymakers, industry experts, ethicists, and civil society. Multidisciplinary perspectives and diverse voices are crucial to navigate the complex ethical challenges of AI.

Who is an "AI actor"?

According to the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), any actor involved in at least one stage of the AI system lifecycle is referred to as an "AI actor". This includes both natural and legal persons, including researchers, programmers, engineers, data scientists, end users, commercial enterprises, academic institutions, and public and private entities.

What kind of ethical concerns are AI systems raising?

AI systems raise new ethical concerns, such as those related to decision-making, employment and labour, social interaction, health care, education, media, access to information, the digital divide, personal data and consumer protection, gender equality, environment, democracy, rule of law, security and policing, dual use, and human rights and fundamental freedoms, such as the right to privacy[150], freedom of speech, and the equality before the law.

Moreover, the potential for AI algorithms to reproduce and reinforce pre-existing prejudices and intensify existing forms of discrimination, prejudice, and stereotyping presents significant ethical challenges. Long-term, AI systems may undermine the added value previously ensured through humans' unique sense of agency and experience, bringing new questions regarding human self-awareness, social, cultural, and environmental interactions, as well as autonomy, agency, value, and dignity.[151]

149 Ibid.

150 A noteworthy document in the area of privacy and data protection issues for the Judiciary is the UNESCO (2022) Guidelines for Judicial Actors on Privacy and Data Protecti on, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381298

151 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455

Activity: Does AI make better decisions than humans? Thinking Ethics of AI

Training participants watch the video and discuss how AI and ethics interact and what the impact of AI is on ethics and human rights.

Source: UNESCO, https://youtu.be/2E7l1hdjHsg

Key Frameworks for AI Ethics

In addition to the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), some other frameworks are briefly presented below:

- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: The IEEE Standards Association has developed a series of documents, including the Ethically Aligned Design framework[152] and the P7000 series of standards[153]. These resources provide a comprehensive approach to AI ethics, covering areas such as transparency, accountability, and the prioritization of human values.[154]
- The European Commission's Ethics Guidelines for Trustworthy AI: The European Commission published guidelines that outline seven key requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination, and societal and environmental well-being.[155]

152 IEEE (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), available at:https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
153 See: https://sagroups.ieee.org/7000/
154 See: https://standards.ieee.org/industry-connections/ec/autonomous-systems/
155 European Commission (2019). Ethics guidelines for trustworthy AI, available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- On November 3, 2017, the Montreal Declaration for Responsible AI was announced at the end of the Forum on the Socially Responsible Development of AI held in Montréal. The Declaration is an example of a co-creation effort that developed a set of guiding principles for the responsible development and deployment of AI for public purposes. This was a collaborative effort involving a series of public consultations and citizen assemblies with over 500 residents, experts, and key stakeholders. With more than 2,200 citizens and over 200 organizations signing the declaration, it advocates for the following principles: Well-being, Privacy and Intimacy, Respect for Autonomy, Responsibility, Democratic Participation, Equity, Solidarity, Diversity and Inclusion, Prudence, and Sustainable Development.[156]
- The Asilomar AI Principles: These principles were developed by a group of AI researchers, policymakers, and thinkers during the Asilomar Conference on Beneficial AI. They cover various ethical aspects, including ensuring AI's broad benefits, long-term safety, technical research leadership, and cooperative orientation.[157]
- The OECD AI Principles prioritise the development of trustworthy AI with a human-centered approach. Crafted with input from a panel of over 50 experts spanning governments, academia, business, civil society, international organizations, the tech community and trade unions, there are five principles centered around values for the responsible and trustworthy implementation of AI, as well as five recommendations for public policy and global collaboration. Their objective is to provide direction to governments, organizations, and individuals in the development and operation of AI systems that prioritise people's well-being, and to ensure that those responsible for their functioning are held accountable.[158]

Table 5 at the beginning of module four gives an overview of the key initiatives on AI regulation, policy, and ethics.

How to operationalize AI Ethics?

Any AI initiative in the Judiciary must adhere to the ethical norms of accountability and openness. The IEEE recommends creating new standards that specify quantifiable, testable degrees of transparency so that systems can be impartially evaluated, and the degree of compliance may be established to sustain transparency.

156 See: https://gouai.cidob.org/resources/montreal-declaration-for-a-responsible-development-of-artificial-intelligence/
157 Future of Life Institute (2017). AI Principles, available at: https://futureoflife.org/open-letter/ai-principles/
158 OECD (2019). Forty-two countries adopt new OECD Principles on Artificial Intelligence, available at: https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm

Yet, due to the intricately linked and layered processes of algorithmic programming, maintaining algorithm transparency is becoming more and more difficult.[159] Codified data ethics principles or conduct codes, ethical impact assessments and privacy impact assessments, ethical training for judicial operators, and ethical review boards are a few examples of ethical review methods that can enable greater transparency and accountability in the use of AI and ADM systems in the justice system.

In general, privacy impact assessments enable organisations and developers to efficiently evaluate the risks posed (ensuring compliance with privacy requirements, identifying mitigation measures, and effectively classifying the impacts of data and algorithm use). It would also be ideal to take a stakeholder-inclusive approach that emphasizes "the proactive inclusion of users." Additionally, the context of data utilization should constantly be considered, necessitating human intervention and occasionally context-specific expertise.[160]

2. What is AI bias?

AI bias is a systematic difference in the treatment of certain objects, people, or groups (e.g., stereotyping, prejudice or favouritism) compared to others by AI algorithms. AI bias can impact data collection and interpretation, system design, and how users engage with a system.[161]

AI systems are far from being neutral pieces of technology. Instead, they can reflect the (un)conscious preferences, priorities, and prejudices of their creators. Biases can arise in many ways in AI systems. Training data and AI models may be biased. Privileged groups may have advantages compared to other groups in AI decisions.

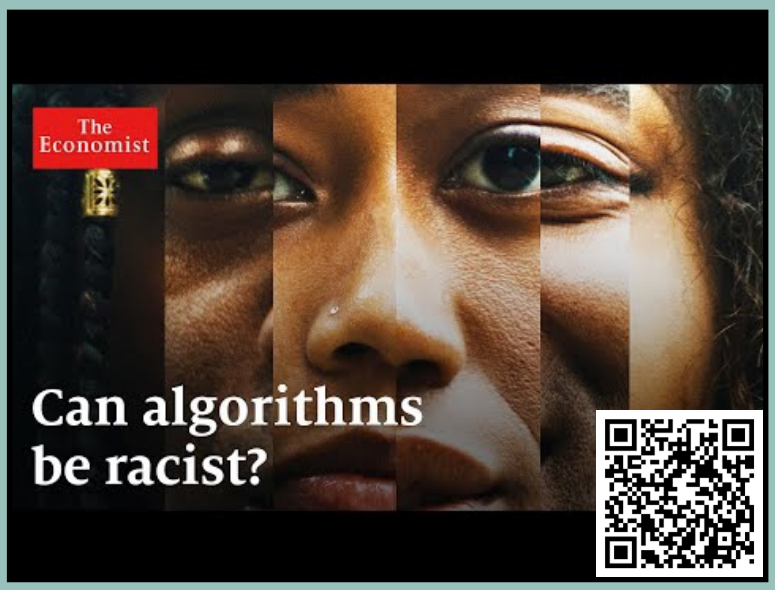
Even when software developers take great care to minimize any influence by their own bias, the data used to train an algorithm can be another significant source of bias. AI systems may reinforce what they have learned from data and increase risks such as racial and gender bias.[162]

Furthermore, even a carefully constructed algorithm must base its judgments on information from an unpredictable and imperfect reality. AI programs are susceptible to making judgement errors in novel situations.[163]

159 See: https://www.ieee.org
160 Morley J., Floridi L., Kinsey L., Elhalal A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, Eng Ethics, 26, 2141–2168, available at: https://doi.org/10.1007/s11948-019-00165-5
161 Goole (2023). Machine Learning Glossary, available at: https://developers.google.com/machine-learning/glossary/
162 Turner Lee N., Resnick P., Barton G (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms, available at: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/
163 Judge Dixon H. B. (2021). Artificial Intelligence: Benefits and Unknown Risks, available at: https://www.americanbar.org/groups/judicial/publications/judges_journal/2021/winter/artificial-intelligence-benefits-and-unknown-risks/

Discussion point

Training participants watch the video and discuss how AI bias has affected them and why it is important to be aware of it in judicial settings



Source: The Economist, https://youtu.be/IzvgEs1wPFQ

Thought experiment: Data driven biases in identifying cats and dogs

Imagine you're creating an AI program to recognize pets. If the algorithm is trained on a million dog images, but only a few thousand cat pictures, it may struggle to accurately identify cats due to a less developed understanding of their appearance. It's worth noting that AI can exhibit bias, as it relies on data and training choices that may be influenced by human biases.

Source: Utrecht University, Unboxing the black box of AI, available at: https://www.uu.nl/en/organisation/in-depth/unboxing-the-black-box-of-ai

Some of the most controversial biases in AI occur in facial recognition technology. A 2016 study conducted in Oakland, California found that despite survey data showing an even distribution of drug use across racial groups, algorithmic predictions of police arrest were concentrated in predominantly African American communities, creating feedback loops that reinforced patterns of systemic bias in the history of police arrests.[164] Algorithms can also introduce racial biases when facial recognition algorithms are trained

164 World Bank WDR 2021. The 2016 study conducted by the Human Rights Data Analysis Group using 2010 and 2011 data from the Oakland police department and other sources compared a mapping of drug use based on survey data from the victims of crime with another based on algorithmic analysis of police arrests. The study showed that biased source data could reinforce and potentially amplify racial bias in law enforcement practices. Data on arrests showed that African- American neighborhoods have on average 200 times more drug arrests than other areas in Oakland.

predominantly on data from Caucasian faces, significantly reducing their accuracy in recognizing other ethnicities.[165] It is concerning that various technologies do not perform accurately for individuals with darker skin.

For instance, a study conducted by Georgia Tech has revealed that driverless cars are more likely to hit people of colour, as the object detection systems they use to identify pedestrians do not work as effectively on individuals with darker skin. These examples highlight the need for more inclusive and unbiased technology that caters to everyone, regardless of their skin colour.[166] The tech industry has been facing a long-standing issue of diversity in its workforce. Mozilla's 2020 Internet Health Report suggests that almost 80% of employees at major tech giants like Apple, Facebook, Google, and Microsoft are male. Furthermore, there has been minimal growth in the representation of Black, LatinX, and Native communities since 2014, which is an alarming concern that needs to be addressed.[167]

165 Hill K. (2020). Wrongfully Accused by an Algorithm." New York Times, available at: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

166 Kenny C. (2021). Artificial Intelligence: Can We Trust Machines to Make Fair Decisions? Data and Computer Scientists, Ecologists, Pathologists, and Legal Scholars Study AI's Biases, disponible en: https://www.ucdavis.edu/curiosity/news/ais-race-and-gender-problem

167 Mozilla (2020). Internet Health Report, available at: https://foundation.mozilla.org/en/insights/internet-health-report/

In depth: Examples of AI bias

Microsoft Tay was created to appeal to individuals between the ages of 18 and 24, and it debuted on social media with a cheery "Hello, world!" (the "o" in "world" was an emoji of the planet Earth). Within twelve hours, however, Tay transformed into a foul-mouthed, racist Holocaust denier who stated that all feminists "should die and burn in hell." Tay, which was swiftly deleted off Twitter, was designed to learn from the actions of other Twitter users, and in this aspect, it was successful. Tay's acceptance of humanity's worst characteristics is an example of algorithmic bias, which occurs when seemingly harmless code adopts the biases of its designers or the data it is fed.

In 2015, Google Photos misidentified several African American users as gorillas, igniting social media outrage. Google's chief social architect and head of infrastructure for Google Assistant, soon announced on Twitter that a team was being assembled to address the issue.

Source: Wired (2017). How to Keep Your AI From Turning Into a Racist Monster, available at: https://www.wired.com/2017/02/keep-ai-turning-racist-monster/; see also: https://www.bbc.com/news/technology-33347866.

AI tools can be biased towards people of colour and minorities

A 2019 US National Institute on Science and Technology (NIST) research of facial recognition technologies, which are often "AI-based," discovered that the algorithms were up to 100 times more likely to yield a false positive for people of colour. For instance, the NIST discovered that "for one-to-many matching, the team saw greater rates of false positives for African American females," a finding that is "especially significant because the repercussions might include wrongful allegations." The error rate for dark-skinned women was 34.7%, but the error rate for light-skinned men was 0.8%, according to a second study conducted by Stanford University and MIT. An assessment of Rekognition, a facial recognition system owned by Amazon and sold to law enforcement, discovered indicators of racial bias and found that the system incorrectly recognized 28 members of the US Congress as convicted offenders. Similarly, AI and algorithmic decision-making systems employed in pre-trial dispositions, sentencing, and prison contexts frequently provide erroneous or biased outcomes that perpetuate existing disparities.

One of the most challenging aspects of AI bias is that AI engineers and developers need not be intentionally racist or sexist. This is a worrisome condition in a time when people increasingly believe that technology is more impartial than they are. As the computer industry develops AI, it runs the risk of incorporating racism and other prejudices into code that will make choices for decades. And because deep learning implies that code, not humans, will write code, the need to eliminate algorithmic bias is even higher.

AI bias can be caused by various reasons, and the below are some definitions and examples of major AI biases:

- **Sample bias due to biased and non-representative training data:** If the rules extracted by the machine learning algorithm from any given set of data are considered legitimate, prejudices and omissions embedded in the example data will be repeated in the predictive model. In other words, if the data used to train the AI model is not representative of the context in which the AI system will be used, the AI system may produce biased outputs. For example, a facial recognition system that has been predominantly developed using photographs of white men, may not be able to identify accurately women or other racial groups. Research shows that in the case of women and people of different racial and cultural backgrounds, these models' levels of accuracy are significantly lower. Another example would be AI systems programmed to identify skin cancer. If the initial dataset is not representative of the population, this method will perform poorly for members of underrepresented groups.[168]
- **Recall bias during the labelling of data:** when the AI solution uses labelled data, the labelling process should be consistent across datasets, otherwise, the result from the model becomes inaccurate. For instance, someone might describe one image of a phone as damaged but another, comparable image as slightly damaged. The dataset will be inconsistent in this situation as there will be two different labels referring to similar and comparable images.
- **Association bias:** It is important to note that even representative data sets reflect historical and societal biases, for example against minorities overly-represented in prison populations or women in less prestigious jobs. The data's 'representativeness' can therefore perpetuate discrimination and inequality, when in fact a consciously adapted dataset that corrects for such social inequalities might produce less discriminatory out-comes from algorithms trained on this basis and then applied to fresh cases (such as when used for informing custodial sentencing or automated scrutinizing of job applications).The best-known association bias is gender bias, such as when the dataset used refers to a group of professions where all the men work as

168 Mozilla (2020). Internet Health Report, available at: https://foundation.mozilla.org/en/insights/internet-health-report/

doctors and all the women as nurses. This does not preclude males from becoming nurses or women from becoming doctors. However, according to the ML model, there are no male nurses or female doctors.

- **Measurement bias:** is caused by faulty measurement by subjects and/or researcher. The source of measurement bias is an inaccuracy made during data collection or measurement. For instance, if the photos captured by a camera used to provide data for an image recognition system are of low quality, this might result in biased findings against certain demographics.[169] Another illustration is human judgment. For instance, a medical diagnostic system can be trained to predict the probability of illness based on proxy measures such as doctor visits rather than real symptoms.[170] Measurement bias can also stem from when the data for certain groups of population is not captured at all because of their existence outside the data-gathering stream. For instance, using mobile phone data as a proxy indicator of the user's ability to repay loans may disadvantage people with limited or no access to mobile phones. Another example would be a situation where an algorithm designed to find candidates for potentially successful jobs may use past success in the workplace as a predictor of future success in the workplace and extract from that information specific favoured recruiting criteria like education and experience. The underlying statistics, however, can be outdated, for instance from a time when minorities or women were underrepresented in the relevant job market or school admittance standards. As a result, the system might disqualify applicants who might outperform the "successful job performer" dataset from the past.[171]
- **Automation bias due to uncritical reliance on AI generated outputs:** A major threat posed by the use of AI systems in the administration of justice is the so-called automation bias, which is the tendency of humans to uncritically consider the solution offered by AI as correct. This can lead to a lack of skepticism towards the information provided by algorithms and a tendency to act automatically on what the algorithm suggests. Detecting automation bias can be difficult as it is often unconscious. One way to detect it is to pay attention to how we rely on the information

169 Hackernoon (2020). 7 Types of Data Bias in Machine Learning, available at: https://hackernoon.com/7-types-of-data-bias-in-machine-learning-ubl3t3w.
170 Data Camp (2022). Different types of AI bias, available at: https://www.datacamp.com/blog/data-demystified-the-different-types-of-ai-bias.
171 Baker J. E., Hobart L. N., Mittelstead M. G. (2021). AI for Judges. A Framework. Center for Security and Emerging Technology, available at: https://www.armfor.uscourts.gov/ConfHandout/2022ConfHandout/Baker2021DecCenterForSecurityAndEmergingTechnology1.pdf

provided by automated systems and to question whether we are being critical of that information or whether we are accepting it without question. It is also important to be aware of our own biases and prejudices and to try to be objective when evaluating the information provided by automated systems. Therefore, the judge's departure from any decision that is assisted or automated should not involve any form of reprisal, sanction, inspection or disciplinary regime. If human supervision and control prevails, the control must be effective (see section on "The human in the loop principle" in Module 1, p. 43-45).

Activity Training participants read the story below and assess the ethical impact of the technology following UNESCO's Ethical Impact Assessment instrument in Annex I (please focus on the parts that deal with fairness, non-discrimination, diversity, and data protection and privacy).

In 2020, JK applied for an international driver's licence at the State Office of Transportation in Hamburg, a northern port city in Germany. She brought all the required paperwork to her appointment, except for a biometric photo since she wanted to take it at the office photo booth. In order to take a biometric photo, she had to place her face in a specific area of the camera, and the picture would only be taken once the face was detected there. JK was not recognized in the State Office of Transportation's photo booth as it appeared that only faces with light skin tones were recognized by the facial recognition software in this photo booth.

JK remembered a staffer saying that there might be a problem with her skin tone. The government printing office was the owner of the photo booth. They said that since the photo booth was equipped with the most recent technology, the issue was not software related. Instead, the office claimed that the lighting in the booth was inadequate and was the cause of the issue. The most recent AI technologies, however, can have weaknesses that result in discriminatory and sexist outcomes, according to a study by Joy Buolamwini and Timnit Gebru.

Source: Algorithm Watch. Automated Decision-Making Systems and Discrimination Understanding causes, recognizing cases, supporting those affected A guidebook for anti-discrimination counselling; Buolamwini J., Gebru T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 81, 77–91, available at: https://proceedings.mlr.press/v81/buolamwini18a.html

Reminder!

AI tools incorporate the policy choices of previous decision makers, and thus the bias from those decisions. Pre-drafted judgement tools, for example, may introduce prejudices, reduce judicial discretion, and fail to address the specific difficulties confronting individuals from marginalized and vulnerable groups. As a result, understanding these technologies and continuing to investigate and evaluate them will ensure that judges can fully participate in the evolution of court operations enabled by AI.

A word of caution - on bias in AI systems leading to discrimination

Even if an AI system appears to be neutral on the surface, its algorithms might lead to discriminating assessments and consequences. Discrimination often can arise from prejudiced practices in the real world which feed into the data used by the AI system.

When data-driven policing technologies are black boxes, it is difficult to analyze the hazards of mistake rates, false positives, limitations in programming capabilities, skewed data, and even faults in source code that influence search results. These black box systems perpetuate vicious cycles of bias.

Predictive policing systems that rely on historical data run the risk of replicating the outcomes of previous discriminatory acts. This can result in "feedback loops," in which each new choice based on prior data generates more data, resulting in marginalized groups being disproportionately suspected and jailed. Predictive algorithms may contribute to biased decision-making and discriminatory consequences depending on how crimes are documented, which crimes are chosen to be included in the study, and which analytical methods are employed.

Though many people believe that police data is neutral, it contains political, social, and other biases. Data from the police department reflects the department's procedures and priorities, as well as local, state, and federal interests, and institutional and individual prejudices. There are no defined procedures for using information gathered during law enforcement operations in the development of AI systems. Further, police practices may have limited openness and supervision.[172]

Many research studies have shown repeatedly that the use of predictive

172 Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, available at: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf. "Increasingly governments are adopting regulations regarding the use of data, such as the European Union's General Data Protection Regulation (GDPR). But these tend to be focussed on corporate use of data and the extent that protections afforded under these regulations extend to LEAs [Law enforcement agents] is less clear", available at: UNESCO (2022). Global toolkit for law enforcement agents: freedom of expression, access to Information and safety of journalists, available at: https://unesdoc.unesco.org/ark:/48223/pf0000383978

algorithms in policing trained on past crime data replicates and amplifies existing systemic biases. Often, this process has little consideration to how different crime reduction initiatives, crime legislation, profiling tendencies, or sentencing biases influence the patterns detected by such algorithms in the data.

Increased public scrutiny of these algorithms has raised questions about how they are developed and implemented; why they are not subjected to greater scrutiny; and whether there are governance mechanisms in place to properly assess their risks, vulnerabilities, and potential for greater societal harm.[173] It has been shown that the deployment of AI tools in the criminal justice system can exacerbate already discriminatory policing practices against minorities.

Activity: The Truth About Algorithms. Training participants watch the video presented by Cathy O'Neil and discuss how and why algorithms are biased. Participants also discuss how algorithmic bias might impact their work.

Source: https://youtu.be/heQzqX35c9A

After watching the video training participants also discuss the following scenario:

Scenario: Algorithmic Bias in Hiring

In the not-so-distant future, a large corporation, let's call it "TechCo," decides to implement an algorithmic hiring system to streamline their recruitment process and make it more efficient. TechCo prides itself on its commitment to diversity

173 Grupo de trabajo de la NACDL sobre vigilancia predictiva (2021). Garbage in, gospel out. How Data-Driven Policing Technologies Entrench Historic Racism and 'Tech-wash' Bias in the Criminal Legal System, disponible en: https://www.nacdl.org/getattachment/eb6a04b2-4887-4a46-a708-dbdaade82125/garbage-in-gospel-out-how-data-driven-policing-technologies-entrench-historic-racism-and-tech-wash-bias-in-the-criminal-legal-system-11162021.pdf

and inclusion, and the leadership believes that using AI-driven hiring tools will help them achieve these goals. They hire a team of data scientists and machine learning engineers to develop the system.

Here's how the scenario unfolds:

1. Data Collection:
   - The team starts by collecting historical data from TechCo's past hiring processes. This dataset includes resumes, interview feedback, and hiring decisions from the last decade.
   - Unfortunately, the historical data reflects some biases that have existed within the company. For instance, there is a disproportionate number of male candidates hired for technical roles, and candidates from certain prestigious universities are favored.
2. Model Training:
   - The data scientists use this historical data to train the algorithm. They aim to identify patterns and criteria that predict successful candidates.
   - Due to the biased historical data, the algorithm starts picking up these biases. For instance, it might learn that candidates from prestigious universities are more likely to be successful, even though this preference is based on historical bias rather than objective merit.
3. Unintended Bias:

   As the algorithm starts processing new job applications, it inadvertently perpetuates the biases present in the training data. Resumes from women, candidates from underrepresented backgrounds, and those from less prestigious schools receive lower scores, leading to their rejection or being pushed to the bottom of the hiring pool.
4. Complaints and Ethical Concerns:
   - Over time, job applicants who feel they were unfairly rejected begin to voice their concerns. They notice a pattern where the algorithm systematically disadvantages certain groups.
   - Civil rights organizations and media outlets catch wind of these issues and start to investigate TechCo's hiring practices, accusing them of algorithmic bias and discrimination.
5. Legal and Reputational Consequences:
   - TechCo faces legal challenges and potential lawsuits for discriminatory hiring practices. They also suffer a significant hit to their reputation, with customers and partners expressing concern about their commitment to diversity and inclusion.
   - The company's leadership realizes the algorithmic bias issue and decides to temporarily halt the use of the hiring algorithm while they investigate the problem.
6. Algorithmic Audit and Corrective Measures:
   - TechCo hires external auditors and data ethicists to assess the algorithm and its impact. The auditors identify the biased data and the flaws in the model.
   - The company takes steps to retrain the algorithm with a more diverse and representative dataset, remove biased features, and implement safeguards against future bias.
7. Rebuilding Trust:

   TechCo apologizes publicly for the algorithmic bias and discrimination. They outline their commitment to rectifying the issue and ensuring fair hiring practices.

The company invests in transparency measures, regularly publishing reports on the performance of their hiring algorithm and seeking external oversight to regain trust.

Several job candidates who believe they have been wronged by the bias embedded in the hiring system lodge complaints. What will you decide and what factors you will take into account when making your decision?

The bias related risks posed by AI and ADM have become pervasive, such as in facial recognition systems in public spaces that enable mass surveillance[174] or in the deployment of heavily biased ADM systems for welfare fraud detection, such as the Dutch SyRI system, discussed in the Box below.[175] AI systems can work in unpredictable ways, and even systems that seem to do "simple" or routine tasks can have unintended and often damaging results. This makes the risks even higher, as shown in the box below that highlights examples of algorithmic bias in the Judiciary.

Case Studies: Examples of algorithmic bias in the Judiciary and Government

COMPAS

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) used by the Judiciary in the United States does not include race or ethnicity as a criterion, yet research has shown that it routinely assigns greater risk scores to black defendants than to white defendants, making them less likely to be freed from custody.[176] There have been instances where prisoners with practically perfect records, like Glen Rodriguez[177], have been denied parole due to an inaccurate COMPAS score, leaving them with little recourse to contest the decision or even find out how it was calculated. A 2016 analysis by ProPublica revealed that COMPAS as used by Florida courts contained racial prejudices. ProPublica examined 7,000 instances and discovered that the score was extraordinarily unreliable in predicting violent crime: just 20% of those expected to commit violent crimes did so. The researchers also discovered that the algorithm was more likely to designate defendants of colour as future criminals than white defendants, and that white defendants were more frequently mislabeled as low risk than defendants of colour.[178] The owner of COMPAS, Northpointe, published a rejoinder, which responded to the ProPublica study and argued that ProPublica's report was "based on faulty statistics and data analysis and failed to show that the COMPAS itself is racially biased, let alone that other risk instruments are biased"[179].

174 Big Brother Watch (2019). UK MASS SURVEILLANCE CHALLENGED IN EUROPE'S HIGHEST HUMAN RIGHTS COURT, available at: https://bigbrotherwatch.org.uk/2019/07/uk-mass-surveillance-challenged-in-europes-highest-human-rights-court/

175 Algorithm Watch (2020). How Dutch activists got an invasive fraud detection algorithm banned, available at: https://algorithmwatch.org/en/syri-netherlands-algorithm/

176 Corbett-Davies S., Pierson E., Feller A., Goel S., Huq A. (2017). Algorithmic decision making and the cost of fairness, available at: https://arxiv.org/pdf/1701.08230.pdf

177 Wexler R. (2017). When a computer program keeps you in jail: How computers are harming criminal justice, available at: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

178 Criswell B. (2020). Algorithms Deciding the Future of Legal Decisions, available at: https://montrealethics.ai/algorithms-deciding-the-future-of-legal-decisions/

179 Angwin J., Larson J., Mattu S., Kirchner L. (2016). Machine Bias, available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; also see: https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder.

The SyRI system

In order to detect social welfare fraud, the Dutch government deployed a system called SyRI, which stands for "system risk indication", to cross-reference residents' personal information from different databases and identify "unlikely citizen profiles" that require further scrutiny. The system functioned as follows: if a government agency (e.g., municipalities, the social security bank, tax authorities) detected fraud with benefits, allowances, or taxes in a certain neighbourhood, it could use SyRI. SyRI helped identify which residents needed to be further investigated for fraud.

This practice was opposed by the Dutch Data Protection Authority and the Council of State, which raised concerns about the right to privacy as well as due process rights, such as the presumption of innocence. Furthermore, the system lacked transparency as its algorithms were not published and it did not undergo a technical audit, and its targeting of disadvantaged neighbourhoods could amount to discrimination based on socioeconomic or migrant status of the residents. What is more, SyRI has been used mostly in low-income neighbourhoods. This exacerbates discrimination and bias if the government exclusively uses SyRI's risk analysis in such neighbourhoods.

In 2020 the court of The Hague ordered[180] the immediate halt of SyRI, whereby it concluded that the legislation establishing SyRI provided insufficient protection against intrusion in private life, owing to disproportionate steps adopted to prevent and punish fraud in the interest of economic well-being. The court concluded that SyRI violated article 8 of the European Convention on Human Rights (ECHR), which protects the right to respect for private and family life.

Source: Algorithm Watch (2020) How Dutch activists got an invasive fraud detection algorithm banned, available at: https://algorithmwatch.org/en/syri-netherlands-algorithm/. See also: https://towardsdatascience.com/fighting-back-on-algorithmic-opacity-30a0c13f0224; https://iapp.org/news/a/digital-welfare-fraud-detection-and-the-dutch-syri-judgment/; https://pace.coe.int/en/files/28715/html

Right to explanation regulation in the EU in the context of ADM

Rules such as the EU General Data Protection Regulation (GDPR)'s "right to explanation" were enacted in response to problems related to AI transparency and accountability.[181] Articles 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR mandate upon data controllers to inform data subjects of the existence of ADM, including profiling, referred to in Article 22(1) and (4) and meaningful information about the logic involved, as well as the significance and the consequences for the data subject.[182]

180 Algorithm Watch (2020). How Dutch activists got an invasive fraud detection algorithm banned, available at: https://algorithmwatch.org/en/syri-netherlands-algorithm/.

181 Casey B., Farhangi A., Vogl R. (2018). Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise, Berkeley Technology Law Journal, 34, available at: https://ssrn.com/abstract=3143325

182 A data subject is an individual who can be identified, either directly or indirectly, through an identifier like a name, ID number, or location data, or through personal factors related to their physical, physiological, genetic, mental, economic, cultural, or social identity. See also: https://academic.oup.com/idpl/article/7/4/233/4762325

Article 22(1) of the GDPR specifies that data subjects are entitled not to be subject to a decision based exclusively on automated processing, including profiling, that creates legal effects concerning them or similarly significantly affects them. Article 22(2) — (4) outlines the limited conditions in which automated decision-making is permissible and outlines certain protections to ensure that data subjects can successfully exercise their rights.[183]

Case Study: Right to explanation legislation in Estonia

The Estonian Unemployment Insurance Act's Section 23(4) enables the Unemployment Insurance Fund to make decisions about the allocation of unemployment benefits to applicants entirely automatically. Applicants are immediately informed that the decision was made automatically, that they have a right to be heard, and that they can file a request for an internal review.

Such practices allow the people whose date has been subjected to automated decision making to understand how the decisions were made and appeal such decisions.

Source: https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report%ef%bf%bc/

AI bias and gender equality

For instance, Automated Gender Recognition (AGR) technologies remove the right to self-identification and infer gender based on acquired data about persons. AGR technologies use information such as a person's legal name and facial features to simplify gender identity to a binary. This lacks a scientific understanding of diverse gender identities.[184] This systematic and technologically reinforced erasure has real-world effects on the fundamental rights of individuals with diverse gender identities and affect the enjoyment of their rights related to social assistance, such housing, work, and healthcare benefits.[185] Moreover, the design of datasets can affect the identity of individuals. A dataset that captures gender as binary, for instance, misgenders individuals with diverse gender identities.[186]

183 Ibid.

184 Sun S. D. (2019). Stop Using Phony Science to Justify Transphobia, available at: https://blogs.scientificamerican.com/voices/stop-using-phony-science-to-justify-transphobia/; see also UN, OHCHR and the human rights of LGBTI people, available at: https://www.ohchr.org/en/sexual-orientation-and-gender-identity. See also, UNESCO (2022). Glossary: Understanding concepts around gender equality and inclusion in education, available at: https://unesdoc.unesco.org/ark:/48223/pf0000380971

185 Leufer D. (2021). Computers are binary, people are not: how AI systems undermine LGBTQ identity, available at: https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/

186 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

AymurAI: Responsible Artificial Intelligence for open and gender-sensitive justice

AymurAI is an initiative to promote open gender-sensitive justice in Latin America. This initiative aims to help criminal court officials and judges that want to promote open data in their criminal courts. AymurAI is a software based on Artificial Intelligence (AI) that semi-automatically identifies important information in judicial rulings and creates open datasets focusing on gender-based violence data. It also has an anonymization tool that detects sensitive information in criminal court rulings and redacts it. "AymurAI" means "harvest" in Quechua. This tool aims to "harvest" data from judicial resolutions in general, with a specific emphasis on cases of gender-based violence. It is "semi-automated" because it does not operate autonomously without human intervention and decision-making. AymurAI helps detect relevant information and streamlines the collection of court sentences, but human validation of the software's findings is crucial to ensure accurate and reliable results.

AymurAI is a desktop application that reads the court resolution, detects relevant information, presents it to the user for validation, and then stores it in a dataset that can be published. The tool uses rules and Named Entity Recognition (NER) to extract essential information from judicial documents. In cases of gender-based violence, the tags can represent the type of violence, location, gender, relationship with the perpetrator, the judge's ruling in that case, and other relevant data. These tags go through a validation process, and once approved, the collected information is structured into an open dataset. All of this is achieved in four simple steps.

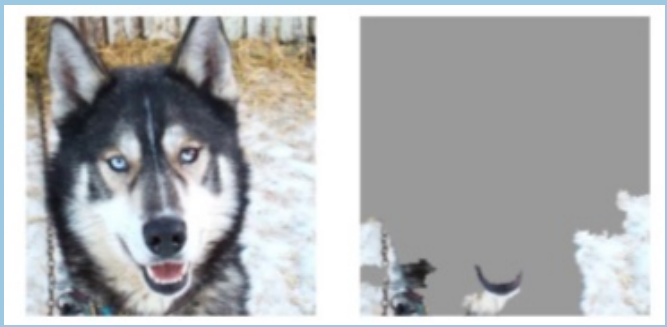
The project arose from the lack of unified data on gender-based violence in Argentina (the only official open database being the one from the Office of Domestic Violence of the Supreme Court of Justice and the reports from the Unique Registry of Gender-Based Violence Cases, which only has data until 2018).While AymurAI can help share information about gender based violence and how it is addressed in different judgements.

AymurAI is currently being implemented at the Criminal Court 10 of the City of Buenos Aires. This Criminal Court, led by Pablo Casas, promotes, designs and enables the application of open justice policies through its public database. This database is maintained by the people who work in the court. The database has around five thousand anonymised legal rulings dated from August 2016 onwards, including many GBV cases. It contains 64 categories with detailed information on each legal ruling, like the type of violence suffered by the victim in each case, in line with the No. 26.485 Argentina Law. The database also includes contextual data (e.g., socio-economic variables

of the people involved in the conflict, whether the defendant has children with the victim, and the phrases used during the aggressions). Employees of Court 10 use different tools to maintain the database. For example, they use a tool to anonymise legal rulings called IA2[187].

Activity: AI can create unanticipated risk that can have life-threatening outcomes. Read the example below and discuss the questions with the participants.

A purposefully flawed algorithm was developed by researchers at University of Washington who categorized photos of husky dogs and wolves. The algorithm exploited the presence or lack of snow to distinguish between domestic huskies and wild wolves. On the training dataset, wolves appeared in the snow more frequently than huskies. Therefore, all images of lupine dogs with snow were classified as wolves by the system. As a result, the AI stood to provide results incorrectly 50% of the time.[188]



Because the pixels that define wolves are those of the snowy backdrop (on the right), a husky (on the left) is mistaken for a wolf. This artifact results from an inadequately represented learning base.

Source: Besse P., Castets-Renard C., Garivier A., Loubes J.— M. (2018). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391288

This example shows It could be fatal if AI systems used in high-stakes fields are trained using inadequately represented learning base.[189] For instance, in the healthcare system data from specific population groups tends to be missing from the data with which ML tools learn, meaning that the tool might work less well for those communities. To illustrate this, a team of U.K. scientists found that almost all eye disease datasets come from patients in North America, Europe, and China, meaning eye disease-diagnosing algorithms are less certain to work well for racial groups from underrepresented countries.[190] Another example is that skin cancer-detecting algorithms tend to be less precise when used on Black patients because ML models are trained chiefly on images of light-skinned patients.[191]

187 https://www.aymurai.info.
188 Pearson D. (2021). AI biopsy dilemma: Wolf or husky, equity or bias?, available at: https://healthexec.com/topics/precision-medicine/ai-biopsy-dilemma-wolf-or-husky-equity-or-bias.
189 Access Now (2018). Human rights in the age of artificial intelligence, available at: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
190 Knight W. (2020). AI Can Help Diagnose Some Illnesses—If Your Country Is Rich, available at: https://www.wired.com/story/ai-diagnose-illnesses-country-rich/
191 Lashbrook A. (2018). AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind, available at: https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

Questions for discussion:

1. What were the main factors used by the system to differentiate between domestic huskies and wild wolves?
2. Were there any flaws in this analysis and why?
3. What would happen if AI decision making processes deployed in the justice systems used similarly flawed algorithms?

3. Why algorithmic transparency and accountability are important in the context of the Judiciary?

The lack of algorithmic transparency is a significant issue at the forefront of discussions on AI and human rights. The deployment of AI systems in the Judiciary is raising concerns about how to thoroughly assess the short- and long-term effects, whose interests do the algorithms serve, and if they are context sensitive to deal with the socio-cultural context in different countries.

This opaqueness of AI systems is alarming. An informed policy debate is impossible without having the ability to understand how AI systems operate. The opacity in how AI systems arrive at their decisions and the difficulty in determining liability for their actions mean that human rights harms can occur when such systems are used.[192]

At the same time, it may also be the case that even when AI-based decisions can be explained, those affected by the decision may not agree with the outcome. In such situations, the affected parties should be entitled to legal recourse. In contrast to the robust procedures that exist in many legal contexts to promote the accountability of human decisions in government—from freedom of information laws to adjudicatory due process protections and appeals procedures—algorithms primarily operate in an accountability-free zone. This section will discuss algorithmic transparency and accountability in the context of judicial operations.

Algorithmic transparency

When dealing with an AI system, transparency refers to how much information is made available to the user. The model's structure, its intended uses, how and when deployment decisions were made, who made those decisions are all part of the transparency, which also includes design decisions and training data.[193]

The users of an AI system deployed in the Judiciary (e.g. plaintiffs and defendants) are often unaware about how the AI system was trained and how

192 Deeks A. (2019). The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. Virginia Public Law and Legal Theory Research Paper No. 2019-51, available at: https://ssrn.com/abstract=3440723

193 Malek Md. A. (2021). Transparency in Predictive Algorithms: A Judicial Perspective, available at: https://doi.org/10.31124/advance.14699937.v2

it takes decisions. Therefore, when it comes to taking legal action against wrong and harmful AI system outputs, it is difficult for those impacted by the use of AI systems to challenge them in the absence of transparency around how the system was designed and how it functions.[194]

The need for algorithmic transparency, include requests to companies to disclose their proprietary algorithms so that they can be reviewed by independent auditors, regulators, or the general public before implementation. However, providing the algorithms or the underlying software code to the public is unlikely, as private companies regard their algorithm as a key proprietary asset and are unwilling to disclose it.

The European Court of Justice has stated that companies cannot state and argue in court that they are not allowed or cannot disclose their algorithms because of Intellectual Property (IP) or trade secret considerations in order to escape from their responsibility to explain AI (under Article 22 GDPR), with the exception of AI that serves a purpose of national security or criminal matters. It has to be noted though that adequate transparency of automated systems is complicated and hard to achieve due to frequent algorithm changes. For instance, Google changes its algorithm hundreds of times per year.[195] Moreover, the risk of manipulating algorithms increases if they are made public.

194 Felzmann H., Fosch-Villaronga E., Lutz C., Tamò-Larrieux A. (2020). Towards Transparency by Design for Artificial Intelligence, Sci Eng Ethics 26, 3333–3361, available at: https://doi.org/10.1007/s11948-020-00276-4

195 https://searchengineland.com/google-seo-news-google-algorithm-updates.

Case study: Algorithmic transparency in practice

- United Kingdom: The UK Central Digital and Data Office and Centre for Data Ethics and Innovation (CDEI) published one of the first national algorithmic transparency guidelines worldwide in 2021. The standard consists of a template that public sector organizations are encouraged to complete for any algorithmic tool that either directly engages the public (such as a chatbot) or meets specific risk-based requirements. The collected information about the AI tools is available in a public register.[196]
- France, the Netherlands, and New Zealand: The three countries have also developed guidance to help public sector officials navigate the responsible use of algorithms. France's Etalab supports government agencies in implementing the legal framework for accountability and transparency of public sector algorithms.[197]
- United States: Several local governments in the United States have implemented bans or temporary halts on using algorithmic technologies, such as facial recognition technologies -FRT, for law enforcement and surveillance. The primary objective of these laws is to address concerns regarding privacy, but there are also significant intersections with algorithmic accountability issues. These bans are typically established through legislation, but some laws have provided limited exceptions to the prohibition, such as third-party information obtained through FRT. For instance, a bill in San Francisco that prohibits using FRT only applies to uses by municipal agencies and excludes usage by federal agencies, such as those in ports and airports.[198]
- Chile: GobLab, an innovation lab within the University of Adolfo Ibanez's School of Government in Santiago, conducted extensive research into the Chilean government's use of algorithms in collaboration with the Chilean Transparency Council. With funding from the Inter-American Development Bank, the group has drafted and proposed regulation that the government is on track to adopt following initial testing of the regulation with various public bodies. The regulation will make Chile the first nation in Latin America to adopt standards on algorithmic transparency.[199]
- City-level initiatives: Algorithmic transparency in the EU has been introduced ex-ante on a local level since October 2020, with the cities of Amsterdam[200], Helsinki[201], and Nantes[202], establishing registers describing the algorithms employed in city administrations. To ensure that AI used by public services is human-centered, the registers indicate, among other things, how data is processed, what dangers are involved, and whether the technologies are subject to human monitoring.[203]

196 Centre for Data Ethics and Innovation (2023). Algorithmic Transparency Recording Standard Hub. gov.uk, available at: https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub
197 Turak H. (2020). Open algorithms: Experiences from France, the Netherlands, and New Zealand. Open Government Partnership, available at: https://www.opengovpartnership.org/stories/open-algorithms-experiences-from-france-the-netherlands-and-new-zealand/.
198 Haataja M, van de Fliert L., Rautio P. (2020). Public AI Registers: Realising AI transparency and civic participation in government use of AI Saidot, available at: https://openresearch.amsterdam/en/page/73074/public-ai-registers
199 Aránguiz Villagrán M. (2022). Algorithmic Audit for Decision-Making or Decision Support Systems. Inter-American Development Bank, available at: http://dx.doi.org/10.18235/0004154
200 See: https://algoritmeregister.amsterdam.nl/en/ai-register
201 See: https://ai.hel.fi/en/ai-register/
202 See: https://data.nantesmetropole.fr/pages/algorithmes_nantes_metropole/
203 Ibid.

Transparency is further complicated by the black box problem of AI systems (discussed in Module 1). Even providing the algorithm's source code may not be enough. It is necessary to explain how the results of an algorithm are generated.[204] One of the most important regulatory goals for the safe and responsible use of algorithms within the public sector is establishing explainability standards.

Case study: Algorithmic transparency from public policymaking perspective: the example of France

In France, the 2016 Law for a Digital Republic stipulates that whenever a public body subjects residents to algorithmic processing, the latter are entitled to be informed of: 1) the degree to which algorithmic processing contributes to decision-making; 2) the data processed; 3) processing parameters; and 4) operations to which such processing is applied. The information should be communicated to an individual upon request in an intelligible language and without infringing on secrets protected by law.

In 2018, when the Constitutional Council was debating a bill to align French data protection law with the GDPR, it ruled that if a public body cannot communicate the operating principles of an algorithm without jeopardizing protected secrets, no decision can be made solely based on such algorithm. Thus, if a public entity bases its decision purely on an algorithm, trade secrecy may not be used to avoid disclosing how the algorithm works.

Source: Décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique, available at: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034194929?r=EILBrO52Ri; also see: https://www.conseil-constitutionnel.fr/decision/2018/2018765DC.htm.

Algorithmic accountability

Algorithmic accountability refers to the ability of those who design, build, procure, or implement the algorithm to be held responsible for their actions and impact according to the policies and laws concerning the use of the algorithm. A governance system holding an actor responsible requires that the actor be able to explain and justify their decisions regarding the algorithm, and face consequences should their actions be against the law.[205]

Accountability by Design

"All AI systems must be designed to facilitate end-to-end answerability and auditability. This requires both responsible humans-in-the-loop across the entire design and implementation chain as well as activity monitoring protocols that enable end-to-end oversight and review."

Source: Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, available at: https://doi.org/10.5281/zenodo.3240529

204 Ibid.

205 Ada Lovelace Institute, AI Now Institute and Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, available at:https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Algorithmic accountability challenges can be related to the fact that the decision maker (e.g., the judge) is not in control of the data sources (data obtained through data brokers or through law enforcement authorities using risk assessment tools). The challenges could also stem from the fact that it is very difficult to translate complex algorithmic concepts (e.g., clustering algorithm results that segment populations based on large numbers of input variables) into human understandable concepts (e.g., racial affiliation). This might result in an inaccurate interpretation of the algorithmic results. Algorithmic accountability challenges can also be triggered by information asymmetries. For instance, the opaqueness of ML algorithms might make it impossible for data subjects to know and understand the results of the Automated Decision Making (ADM) process or to be even aware that they have been subjected to ADM. Also, problems might occur in the implementation stage when adversarial data is injected into the system to fool it into making errors. Please refer to Module 1 talks about cybersecurity issues.[206]

4. Spotlight on biometric identification, facial recognition technology, and deepfakes

The adoption of high-risk technologies, such as facial recognition and biometric identification, presents aggravated challenges to policymakers and regulators worldwide. Human rights NGOs have also called out the lack of proper privacy protections in many national biometric identity systems, where accessing welfare benefits and other government services was found to be contingent upon registration with the system.[207]

In this vein, the UN General Assembly Resolution on the right to privacy in the digital age (2020) has referred to "hacking and the unlawful use of biometric technologies," as "highly intrusive acts that violate the right to privacy" that interfere with freedom of expression and opinion, peaceful assembly and association, and the freedom of religion or belief, and "may contradict the tenets of a democratic society, including when undertaken extraterritorially or on a mass scale."[208]

Moreover, a 2021 United Nations High Commissioner for Human Rights Report "The right to privacy in the digital age" has called for a moratorium on the use of facial recognition technologies in public spaces, until

206 European Parliament (2019). A governance framework for algorithmic accountability and transparency, available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

207 https://www.ohchr.org/Documents/Issues/Poverty/DigitalTechnology/AmnestyInternational.pdf

208 UN General Assembly (2020). The right to privacy in the digital age, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/371/75/PDF/N2037175.pdf?OpenElement

governments can show that there are no substantial issues related to accuracy or discriminatory impacts and that these technologies comply with robust privacy and data protection standards (see infra, p.160).[209]

Biometric recognition is based on comparing a person's digital representation of their face, fingerprint, iris, voice, or movement with other similar representations stored in a database. Based on this, the system decides the probability if the individual is indeed the person to be identified. Authorities across the world are increasingly using remote real-time facial recognition, as a form of biometric recognition.[210]

The UN High Commissioner for Human Rights has indicated that "real-time biometric recognition raises serious concerns under international human rights law".[211] Some of these concerns mirror issues with predictive techniques, such as the likelihood of incorrect identification of persons and disproportionate effects on members of certain (most often marginalized) groups.[212] Individuals can be profiled using facial recognition technology based on their race, ethnicity, national origin, gender/sex, and other traits.[213]

Remote biometric recognition is associated with significant interference with the right to privacy. A person's biometric information is one of the key aspects of their personality since it exposes distinctive qualities that set them apart from other people.[214] Remote biometric recognition enables government authorities' ability to systematically identify and track individuals in public spaces, and this can have a negative impact on the exercise of the rights to free expression, peaceful assembly, and association, and freedom of movement (see infra, p.149).[215]

209 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

210 Ibid.

211 UN Human Rights Council (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement

212 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx.

213 UN Human Rights Council (2020). Racial discrimination and emerging digital technologies: a human rights analysis, paras. 39–40, available at: https://www.ohchr.org/sites/default/files/HRBodies/HRC/RegularSessions/Session44/Documents/A_HRC_44_57_AdvanceEditedVersion.docx

214 UN Human Rights Council (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, para. 33, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement. See also European Court of Human Rights, Reklos and Davourlis v. Greece, Application No. 1234/05, Judgment of 15 April 2009, para. 40.

215 See: European Data Protection Board and European Data Protection Supervisor (2021). Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), para. 30, available at: https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf; UN Human Rights Council (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement; UN Human Rights Council (2019). Surveillance and human rights, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/148/76/PDF/G1914876.pdf?OpenElement

Case Studies

The GDPR and biometric data

The EU's GDPR limits biometric data processing to a certain extent. Only when data is connected to a specific individual, it becomes personal data and thus covered by this Regulation. According to the GDPR, biometric data is "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person." Thus, if biometric recognition is not directed at identifying (but rather at categorization, profiling, or affect recognition), it may not fall under the GDPR definition.

According to GDPR recital 51 "the processing of photographs [is considered] biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person".

Source: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

The case of Clearview AI

The Data Protection Authority from the German State of Hamburg decided Clearview AI unlawfully processed biometric data obtained and made available as a service. Further, there was no valid legal basis for data processing. The court noted that Clearview AI has processed biometric data (under Article 4(14) GDPR), as it "uses a specially developed mathematical procedure to generate a unique hash value of the data subject which enables identification." The litigation was initiated by a data subject complaint since the data subject had not provided consent for his biometric data processing. The Data Protection Authority determined that even though Clearview AI was not established in the EU, it was subject to the GDPR through the monitoring of online activity of data subjects (Article 3(2)(b) GDPR), as it "does not offer a snapshot [of individuals], but evidently also archives sources over a period of time." Clearview AI was ordered to delete all the complainant's personal data.

Source: Future of Privacy Forum (2022). GDPR and the AI Act interplay: Lessons from FPF's ADM Case Law Report, disponible en: https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report

Facial recognition technologies use digital images to identify and validate human faces. These technologies function by identifying face features in a source image and comparing them across a dataset. Facial recognition technologies have a wide range of uses, although they are most commonly employed for security purposes, such as policing and national security activities (e.g., counter terrorism). Advances in AI have improved the capacity and sophistication of these technologies in recent years, making them a standard component of consumer goods such as mobile phones, which allows users to' sign in' using their faces.[216]

216 Hill D., O'Connor C. D., Slane A. (2022). Police use of facial recognition technology: The potential for engaging the public through co-constructed policy-making, International Journal of Police Science & Management, 24(3), 325–335, available at: https://doi.org/10.1177/14613557221089558

Facial recognition technologies controversies in the private sector

Several companies—including Microsoft and IBM—have been criticized for rolling out facial recognition software that is more accurate for some demographics than others. Specifically, these systems tend to accurately identify fair-skinned men far more often than they identify darker-skinned women.

Similarly, controversy arose when Google's automatic photo-tagging software identified many pictures of African Americans as 'gorilla' or 'monkey'. The cause of these errors is likely to lie in the development of the algorithmic models. The models were trained with datasets of photos of predominantly people of Caucasian origin, and thus had not been trained with sufficient data to identify non-white people, particularly women. The work of Joy Buolamwini, a computer scientist at MIT and founder of the Algorithmic Justice League has prompted multiple companies to release statements addressing criticisms and reform their models.

Source: https://www.poetofcode.com/

In November 2021, Meta announced it was "shutting down the Facial Recognition system on Facebook" citing unclear rules from regulators. Likewise, IBM is to stop offering its facial recognition software for certain activities including mass surveillance.

Source: UK Government (2022). Policy paper Establishing a pro-innovation approach to regulating AI, available at: https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement

The use of biometric identification and facial recognition technology in judicial operations can become a Pandora's box for different types of bias, such as those based on race and gender. The case of ImageNet data can be used as an illustrative example. This is a key dataset for the development of computer vision applications, which contains more than 45% of images from the US, compared to just 3% from China and India combined. This lack of diversity contributes to the shortcomings of image recognition algorithms, which interpret Asian eyes as always blinking, label a picture of a traditional US bride dressed in white as "bride," "dress," "woman," and "wedding," but label a picture of an Indian bride as "performance art" and "costume," and misidentify the gender/sex of darker-skinned women with a 35% error rate while misidentifying the gender/sex of lighter-skinned men with a 0% error rate.[217]

217 European Parliament (2019). A governance framework for algorithmic accountability and transparency, available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

While AI-enabled mass surveillance through facial recognition involves the collection, storage, and processing of personal (biometric) data (our faces), it also has an impact on our privacy, identity, and autonomy by opening up the possibility of being observed, tracked, and recognized.[218] People may feel pressured to adhere to a particular standard because of the psychological "chilling" effect, altering the balance of power between the person and the government or private company using facial recognition technology.

While facial recognition may have a more pronounced effect on the right to privacy and psychological integrity, one could argue that digital tracking of all aspects of human lives (via location data, IoT data from smartwatches, health trackers, smart speakers, thermostats, vehicles, etc.) could have a similar impact. The heart rate, body temperature, and other types of AI-driven biometric recognition measure or even forecast our behaviour, mental state, and emotions. This can have severe impact on the right to privacy in the online environment.[219]

In depth: Facial recognition systems can mis-identify gender

AI systems for "gendering" individuals in public settings are not futuristic; they are already in use around the globe. In Sao Paulo, Brazil, the Brazilian Institute for Consumer Protection (IDEC) challenged the installation and use of smart billboards that claim to anticipate the emotion, age, and gender of metro riders to provide them with "better adverts".[220]

218 CAHAI Secretariat (2020). Towards Regulation of AI Systems. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law, Council of Europe Study, DGI/2020/16, available at: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.
219 Ibid.
220 See: https://idec.org.br/sites/default/files/acp_viaquatro.pdf

Activity: Training participants watch the video and discuss the societal implications of AI and facial recognition technologies. Participants also discuss how these technologies might impact their work. How do facial recognition technologies impact human rights? What groups are the most vulnerable and susceptible to human rights violations by facial recognition technologies?

Melbourne-based researchers asked human volunteers to judge thousands of photos for the same characteristics and then used that dataset to create the Biometric Mirror. The Biometric Mirror uses AI to analyze a person's face by scanning it, and later displays 14 traits about them, such as their age, race, and perceived level of attractiveness. It uses an open dataset of thousands of facial and crowdsourced evaluations. However, this analysis is often false because the AI generates the analysis based on subjective and biased information provided by initial human volunteers.[221]

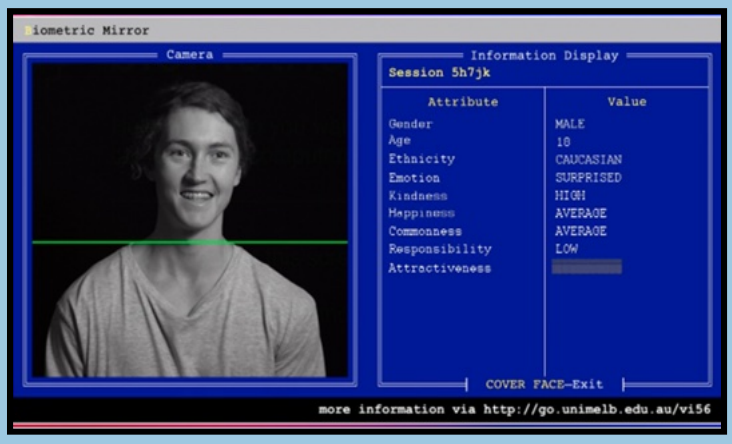


Photo credit: Sarah Fisher/University of Melbourne

Link to video: https://youtu.be/fb_sfhT0mrg

221 Houser K. (2018). Biased AI biometric mirror, available at: https://futurism.com/the-byte/biased-ai-biometric-mirror.

Deepfakes

One particularly dangerous AI technology that impacts human rights is deepfakes. A deepfake is any form of media (video, audio, or other) that has been altered or entirely or partially created from scratch.[222] Machines can learn how to do tasks by looking at examples using neural networks. There are several technologies that may be applied to this, but the most popular one is based on Generative Adversarial Networks (GAN) and Diffusion Models.[223]

222 Van der Sloot B., Wagensveld Y. (2022). Deepfakes: regulatory challenges for the synthetic society. Computer Law & Security Review, available at: https://www.sciencedirect.com/science/article/pii/S0267364922000632, available at: https://doi.org/10.1016/j.clsr.2022.105716

223 Ibid.

Generative Adversarial Networks (GAN)

GANs are an unsupervised approach of deep learning that can generate hyper-realistic material. GANs are used for generating realistic images or image datasets, performing text-to-image and image-to-text translations, ageing faces, and making emojis. GANs use two neural networks: a generator that generates new instances and a discriminator that seeks to differentiate these fake, frequently low-quality or unrealistic images from the real image data input into the AI system. Through this interaction, the generator learns to produce increasingly convincing and high-quality images, which finally deceive the discriminator into believing they are part of the actual image data.[224]

Diffusion models

Diffusion models are generative models that are more advanced than GANs on image synthesis. Most recently, Diffusion Models were used in DALL-E 2, OpenAI's image generation model and Google's Imagen.[225] The public access to DALL-E is controlled via an extensive waiting list and a paywall after a several prompts, while Google's Imagen is off-limits to the public. DALL-E's output is filtered, which makes it difficult to generate images that contain violence, nudity, or realistic faces.[226]

However, the new text-to-image program named Stable Diffusion, developed by Stability AI[227], offers open-source, unfiltered image generation, free to use for anyone. Below is an image created by Stable Diffusion that was created using the exact text "Photo of Bernie Sanders in Mad Max Fury Road (2015), explosions, white hair, goggles, ragged clothes, detailed symmetrical facial features, dramatic lighting."[228]

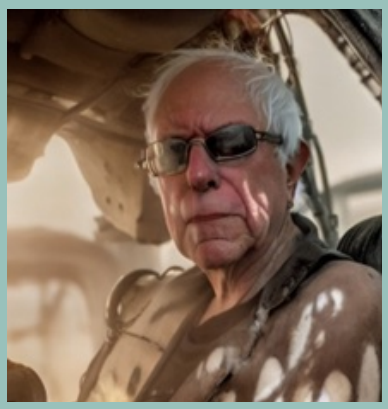


Image: Reddit / Licovoda

Source: The Verge (2022). Anyone can use this AI art generator — that's the risk https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data

As already indicated, in 2023, Getty Images filed a copyright infringement lawsuit against Stability AI in the US, saying that the firm copied 12 million images 'without permission... or compensation' to train its AI model.

224 AAAS. Artificial Intelligence and the Courts: Materials for Judges, available at: https://www.aaas.org/ai2/projects/law/judicialpapers
225 O'Connor R. (2022). Introduction to Diffusion Models for Machine Learning, available at: https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/.
226 See: https://labs.openai.com/policies/content-policy
227 See: https://stability.ai/
228 Vincent J. (2022). Anyone can use this AI art generator — that's the risk https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data

The real problem associated with deepfakes is how simple it is to generate a whole ecosystem of false information. A fake video, fake websites that host the video and generate disinformation and misinformation about what is displayed in the video, fake Twitter accounts that link to the video, fake accounts on discussion forums that discuss the content of the video, fake Instagram accounts that generate memes of the fake video. An environment of deceit that is multi-layered and complex will be exceedingly difficult to penetrate and provide trustworthy information.[229]

Activity: participants watch the video and discuss how deepfakes could affect the work of judicial operators.

Link to video: https://youtu.be/oxXpB9pSETo

Deepfakes and the whole falsified ecosystem that they create endanger the rights to a fair trial, effective remedy, and presumption of innocence (see infra, p.136). They could be used as fake evidence in courts. Parties can always argue that the evidence presented against them is false and contrived, and trials will take longer. Deepfakes also raise the possibility that a judge will mistakenly accept fabricated evidence as dependable.[230] Therefore, the judicial sector should start investing in digital tools that facilitate forensic evaluation of video and audio evidence to ascertain that the evidence has not been generated by GANs and Variational AutoEncoders. On the other hand, AI has the potential to verify the authenticity of digital evidence by detecting fake algorithms or manipulated data. Using AI to analyze an image or video could determine if it has been manipulated in some way. However, this is still a developing area of research.

229 Ibid.
230 Ibid.

These group activities are intended to encourage the training participants to discuss various legal and ethical challenges of AI deployment in the Judiciary.

Activity 1

Training participants read through "State v. Loomis" in Module 4 and answer the following question: Do you think it is appropriate that the court permitted an algorithm, into which players in the legal system have limited visibility, to play even a minor role in depriving a person of their liberty? Please assess the ethical impact of this decision following UNESCO's Ethical Impact Assessment instrument in Annex I [please focus on the parts that deal with fairness, non-discrimination, diversity, and data protection and privacy].

Activity 2

Training participants go through the material on diffusion models presented above and also read the following article: https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion.

Participants then discuss the legal implications of the lawsuit brought by Getty Images against Stability AI. The lawsuit will rely on the interpretation of the US fair use doctrine, which allows unauthorised use of copyrighted works in certain circumstances. The notion of "transformative use" may also be a significant aspect. Is Stable Diffusion's output sufficiently distinct from its training data? Recent study has revealed that the program memorizes some of its training images and can repeat them nearly identically, although in a relatively limited number of instances.

The training participants discuss how AI development and deployment impacts copyright rules in their own jurisdictions.

Activity 3

Discuss the legal and ethical implications behind this case.

Australian Case - Victorian Mayor suiting ChatGPT

A Victorian mayor, Brian Hood, is preparing to sue OpenAI if it doesn't correct ChatGPT's false claims that he had served time in prison for bribery. Hood's lawyers sent a letter of concern to OpenAI on March 21, giving them 28 days to fix the errors, but OpenAI has yet to respond. The false claims were related to a foreign bribery scandal involving a subsidiary of the Reserve Bank of Australia in the early 2000s, but Hood was never charged with a crime .[231]

231 Byron K. (2023). Victorian mayor readies defamation lawsuit over ChatGPT content, available at: https://www.afr.com/technology/vinoctorian-mayor-readies-defamation-lawsuit-over-chatgpt-content-20230405-p5cyh5

Activity 4

Training participants read the text below on how facial recognition technologies can invade the right to privacy and watch the videos. Then, the training participants discuss how facial recognition technologies, and their risks can be litigated under their national data protection and privacy laws.

In May of 2020, the American Civil Liberties Union (A.C.L.U.) filed a lawsuit[232] on behalf of organizations representing domestic abuse victims, illegal immigrants, and sex workers. The organization accused Clearview, a technology firm that develops facial recognition technology, of breaking Illinois's Biometric Information Privacy Act (BIPA)[233], a state statute that prevents commercial businesses from exploiting citizens' physical identifiers, including computational mapping of their faces, without consent.[234]

The complaint was filed in Illinois state court in Chicago after the New York Times disclosed in January 2020 that Clearview was developing a biometric identifier-based tracking and surveillance system. Facial recognition technology has enabled Clearview to acquire more than three billion faceprints from web photographs.[235]

Access to this information has been provided by Clearview to private corporations, affluent people, and federal, state, and local police enforcement organizations. The business asserts that using this large database, it can instantly identify persons with unmatched precision, enabling extensive clandestine and remote surveillance of Americans.[236]

BIPA mandates that businesses that collect, capture, or get a biometric identifier from an Illinois resident, such as a fingerprint, faceprint, or iris scan, must first tell the subject and obtain their written consent. This is due to the fact that the forced acquisition of immutable biometric identifiers poses more dangers to an individual's security, privacy, and safety than the capture of other identifiers, such as names and addresses. And recording a person's faceprint — comparable to establishing their DNA profile from genetic material inevitably shed on a water bottle, but distinct from the publication or transmission of a photograph — is behavior, not speech, and is thus legitimately governed by the law. Clearview did not comply with BIPA, depriving privacy rights to several Illinois citizens.[237]

232 Alba D. (2020). A.C.L.U. Accuses Clearview AI of Privacy 'Nightmare Scenario', available at: https://www.nytimes.com/2020/05/28/technology/clearview-ai-privacy-lawsuit.html.
233 See: https://www.aclu-il.org/en/campaigns/biometric-information-privacy-act-bipa
234 Mac R., Hill K. (2022). Clearview AI settles suit and agrees to limit sales of facial recognition database. The facial recognition software maker is largely prohibited from selling its database of photos to private companies, available at: https://www.nytimes.com/2022/05/09/technology/clearview-ai-suit.html
235 Ibid.
236 Ibid.
237 ACLU (2022). ACLU v. Clearview AI, available at: https://www.aclu.org/cases/aclu-v-clearview-ai

The lawsuit was the first to focus on the harm that Clearview's technology would cause to survivors of domestic and sexual abuse, undocumented immigrants, communities of color, and members of other vulnerable populations. The plaintiff organizations' members, clients, and program participants have been exposed to face printing by Clearview without their consent and stand to suffer some of the most severe effects of Clearview's unparalleled monitoring program.[238]

On May 11, 2022, after the parties negotiated a settlement agreement, the court approved a consent order dismissing this matter. The fundamental element of the settlement bans Clearview's operations not only in Illinois, but throughout the entire United States, permanently prohibiting Clearview from making its faceprint database accessible to private organizations. In addition, the corporation is prohibited for five years from selling access to its database to any agency in Illinois, including state and municipal authorities.[239]

Source: https://youtu.be/s44EFtBoRxY

Source: https://youtu.be/cc0dqW2HCRc

238 Ibid.
239 ACLU, EXIBIT 2. signed settlement agreement, available at: https://www.aclu.org/legal-document/exhibit-2-signed-settlement-agreement

Activity 5

Training participants explore a hypothetical court case involving AI bias and answer how they would decide the case if it was tried in their jurisdiction.

Hypothetical Case Title: Smith v. AI Financial Services

Background: John Smith, an African American, has filed a lawsuit against AI Financial Services, a major lending institution, alleging racial bias in the company's automated loan approval system. He claims that the AI system unfairly denied his mortgage application, leading to financial and emotional distress.

Case Details:

1. **Plaintiff's Argument:** John Smith argues that the AI loan approval system used by AI Financial Services disproportionately denies loans to African Americans, as evidenced by data showing a significant disparity in loan approval rates between racial groups.
2. **Defendant's Response:** AI Financial Services defends its AI system, asserting that it relies on objective financial criteria and does not consider race as a factor in loan decisions. They argue that any disparities in loan approvals are due to differences in applicants' financial histories and creditworthiness.

AI System Examination: During the trial, both parties bring in expert witnesses to examine the AI system:

1. **Plaintiff's Expert:** An AI ethics expert testifies that the AI system's training data had inherent racial bias, which influenced its decision-making. They present evidence of similar cases where AI systems have exhibited discriminatory behavior.
2. **Defendant's Expert:** The defendant's AI expert argues that the AI system was designed to be race-neutral and that any bias in the training data was unintentional. They highlight the rigorous testing and validation processes the AI underwent before deployment.

Court's Role: The judge must determine whether AI bias played a role in John Smith's loan denial and, if so, whether AI Financial Services is liable for discrimination. Key considerations include:

1. **AI System Transparency:** The court evaluates the transparency of the AI system's decision-making process and whether the defendant adequately disclosed its use of AI to loan applicants.
2. **Intent vs. Impact:** The judge distinguishes between intentional discrimination and disparate impact resulting from AI bias, which may still be illegal under anti-discrimination laws.

3. **Mitigation Efforts:** The court examines whether AI Financial Services took reasonable steps to mitigate bias in its AI system and whether it promptly addressed any identified issues.

Outcome: The court finds in favor of John Smith, ruling that the AI system used by AI Financial Services exhibited bias that resulted in disparate impact on African American applicants. The judgment includes financial compensation for John Smith and an injunction requiring AI Financial Services to review and revise its AI algorithms to ensure compliance with anti-discrimination laws.

This hypothetical case highlights the complex legal issues surrounding AI bias in lending and the importance of transparency, fairness, and accountability in the use of AI systems, especially when they impact individuals' rights and access to financial services.

6. Resources

1. Alang N. (2017). Turns Out Algorithms are Racist, available at: https://newrepublic.com/article/144644/turns-algorithms-racist/
2. Angwin J., Larson J., Mattu S., Kirchner L. (2016). Machine bias,, available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Buolamwini J., Gebru T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 81, 77–91, available at: https://proceedings.mlr.press/v81/buolamwini18a.html
4. Commission Nationale de l'Informatique et des Libertés (2022). Asking the right questions before using an artificial intelligence system, available at: https://www.cnil.fr/en/asking-right-questions-using-artificial-intelligence-system
5. Edwards L., Veale M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For, 16 Duke Law & Technology Review, 18, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855
6. European Parliamentary Research Service (2019). A governance framework for algorithmic accountability and transparency, available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf
7. Green B., Chen Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments, Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 90–99, available at: https://doi.org/10.1145/3287560.3287563
8. Hart R. (2017). If you're not a white male, artificial intelligence's use in healthcare could be dangerous, available at: https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous
9. Kleinberg J. , Lakkaraju H., Leskovec J., Ludwig J., Mullainathan S. (2017). Human Decisions and Machine Predictions, available at: https://www.cs.cornell.edu/home/kleinber/w23180.pdf
10. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, available at: https://doi.org/10.5281/zenodo.3240529
11. UTS Human Technology Institute report (2022). outlining a Model Law for facial recognition: https://www.uts.edu.au/human-technology-institute/projects/facial-recognition-technology-towards-model-law
12. Whittlestone J., Nyrup R., Alexandrova A., Cave S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions, Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), Association for Computing Machinery, 195–200, available at: https://doi.org/10.1145/3306618.3314289

Module 4

Human Rights and AI

Module four will give an in-depth analysis of some human rights impacted by AI, such as (i) the right to access to court, fair trial, and due process, (ii) effective remedy, (iii) rights to protection against discrimination, (iv) freedom of expression, (v) right to privacy and data protection, and (vi) access to information. The Module also gives an overview of key governance approaches to AI: risk based, and human rights based.

What will you learn?

After completion of this module, the participants will be able to:

- Understand and explain instances of possible human rights violations through the use of ADM and AI : (i) the right to access to court, fair trial, and due process, (ii) effective remedy, and (iii) the right to protection against discrimination, (iv) freedom of expression, (v) right to privacy and data protection, and (vi) access to information.
- Understand the key governance approaches to AI: risk based and human rights based.

There is a strong correlation between democracy, the rule of law, and human rights. Robust and accountable democratic institutions, inclusive and transparent decision-making processes, and an independent and impartial Judiciary that upholds the rule of law are prerequisites for upholding human rights.

Human rights are the fundamental freedoms and rights that every person has from birth until death. Human rights uphold and defend every person's inalienable dignity regardless of their race, ethnicity, gender, age, sexual orientation, class, religion, level of disability, language, nationality, or any other attribute. Governments are required to uphold, defend, and fulfil human rights. Individuals are entitled to legal remedies that provide for the redress of any human rights breaches.

The International Bill of Rights[240] represents a body of international human rights law that includes nine major human rights treaties; regional rights instruments in the Americas, Africa, and Europe; it has been incorporated in national constitutions and national laws; and customary and case law.[241]

Non — binding intergovernmental instruments such as the UN Guiding Principles on Business and Human Rights[242] have also addressed the issue of responsibility of private sector stakeholders in the context of human rights.

Human rights offer a set of global basic standards based on principles such as equality, autonomy, and human dignity. These principles and the accompanying legal framework impose legally binding obligations on nations to respect, defend, and uphold human rights.

International human rights law requires nation states to provide for an effective remedy where an individual suffers a human rights violation. Effective remedies comprise judicial and administrative remedies, such as ordering compensation or an apology, and preventive measures that may include changes to law, policy, and practice. Human rights obligations also require states to put in place effective mechanisms to prevent human rights being encroached upon.[243]

240 Comprised of the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights.

241 Baluarte D. C., De Vos C. M. (2010). From Judgment to Justice: Implementing International and Regional Human Rights Decisions, Open Society Justice Initiative, Open Society Foundations: New York, available at: https://www.justiceinitiative.org/uploads/62da1d98-699f-407e-86ac-75294725a539/from-judgment-to-justice-20101122.pdf

242 UN Human Rights Council (2011). Guiding Principles on Business and Human Rights, available at: https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

243 ICCPR art 2(3) requires each State Party to ensure a person whose Covenant rights have been violated has an effective remedy, and that this remedy will be enforced. See also: UN Human Rights Committee (2004). The Nature of the General Legal Obligation Imposed on States Parties to the Covenant, available at: https://www.refworld.org/docid/478b26ae2.html

The international human rights law framework is an established means for ensuring the protection of rights in general and in the digital environment, including the rights to equality and non-discrimination. Its nature as an actionable set of standards lends itself especially well to technologies that transcend national boundaries, such as AI. A human rights-based approach provides normative guidance to AI developers to uphold human dignity, regardless of jurisdiction.

Human rights law can inform the development of technical and policy safeguards in AI deployment. In this vein, in 2019, the Human Rights Council (HRC) passed the first resolution (41/11) on "New and emerging digital technologies and human rights".[244] The resolution acknowledges the need to better address the entire spectrum of human rights implications of new technologies to remain relevant in the digital age.

In 2021, the Council adopted resolution 47/23, emphasizing the significance of a human rights-based approach to developing and deploying innovative digital technologies. The resolution notes that new technologies have the potential to offer multiple opportunities to advance human rights by positively contributing to the building of democratic institutions and the resilience of civil society, as well as the achievement of the Sustainable Development Goals (SDGs). Human rights advocates and technology developers, as well as governments, must remain nimble in tackling the human rights concerns posed by AI, using protections and instruments based on existing human rights norms and frameworks.[245]

For AI to benefit the public good, its design and implementation must, at minimum, avoid harming fundamental human values guaranteed by international human rights law, which provides a robust framework for the protection of these values. AI, if adequate safeguards are implemented, could also be a key enabler in enhancing and promoting human rights.

How can AI assist in the protection and fulfilment of human rights?

AI systems have numerous applications that can help in the fulfilment of human rights. For example, AI systems are used to analyse patterns in food scarcity to combat hunger, improve medical diagnosis and treatment, or make health services more accessible.

Module two gave an overview of how AI can assist judicial operators through e-discovery and document review, predictive analytics and ADM support, risk assessment tools, dispute resolution, generative

244 UN Human Rights Council (2019). New and emerging digital technologies and human rights, available at: https://digitallibrary.un.org/record/3834165

245 DiPLO (2022). Promoting and Protecting Human Rights in the Digital Era, available at: https://www.diplomacy.edu/event/promoting-and-protecting-human-rights-in-the-digital-era/

AI, language recognition and analytics, and digital file and case management. The Judiciary, as a public institution, is held to a higher standard when it comes to behaviour of judicial operators, and judges in particular, towards individuals and society. This has been reflected in the rule of law principles such as justification, proportionality, and equality. On the one hand, AI can increase the efficiency of judicial operators, on the other hand, it can also erode the procedural legitimacy of and trust in democratic institutions and the authority of the law.

Without proper guardrails, AI could also encroach on human rights

For instance, undetected bias might be present in ML algorithms that predict recidivism. Or AI deployment could be used to limit people's freedom of expression or their ability to engage in political activity or to identify political dissidents[246]. AI could also harm human rights in situations where there is use of poor-quality training data, system design or complex interactions between the AI system and its environment. One such example is algorithmic exacerbation of hate speech or incitement of online violence. Another example is the amplification of disinformation and misinformation, which could impact the right to participate in political and public affairs, especially during elections. The likely scale and impact of harm will be linked to the scale and potential impact of decisions by any specific AI system. At the same time, it is important to note that AI can be used to identify hate speech and help with taking down content related to promotion of terrorism.

Numerous applications of AI have the potential to directly affect the equality of access to fundamental rights, including the right to privacy and the protection of personal information, the right to access to justice and the right to a fair trial, particularly regarding the presumption of innocence and the burden of proof, the right to employment, education, housing, and health, as well as the right to public services and welfare. If not accompanied by adequate safeguards against bias, AI technologies might contribute to disproportionately deny access to rights to women, minorities, and those who are already the most vulnerable and marginalized.[247]

For example, the use of biometric or facial recognition systems in public spaces might enable mass surveillance encroaching on human rights.[248] According to the AI Now Institute 'Regulating Biometrics' report (2020)[249] facial recognition technology is not an adequate identification

246 UN General Assembly (2018). Promotion and protection of the right to freedom of opinion and expression. Note by the Secretary-General, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement

247 Council of Europe (2019). Preventing discrimination caused by the use of artificial intelligence, available at: https://pace.coe.int/en/files/28809.

248 Human Rights Watch (2020). Argentina: Child Suspects' Private Data Published Online, available at: https://www.hrw.org/news/2020/10/09/argentina-child-suspects-private-data-published-online

249 Kak A. (2020). Regulating Biometrics. Global Approaches and Urgent Questions, available at: https://ainowinstitute.org/publication/regulating-biometrics-global-approaches-and-open-questions

replacement for fingerprints. Facial recognition technologies show poor performance results and high error rates for 'black women, gender minorities, young and old people, members of the disabled community, and manual labourers'.[250]

Often, deployment of AI by law enforcement agencies might encroach on due process and equal protection rights. For instance, if the AI system is used for DNA testing that involves processing of sensitive health data, and criminal justice risk assessments that might be biased towards certain populations based on gender/sex, race, ethnicity etc.

Reminder!

As we have seen, predictive policing or facial recognition tools cannot be a predetermination of guilt or evidence sufficient to rebut the presumption of innocence. A statistical prediction cannot be a cause for arrest or, under common law, reasonable suspicion, or a step higher, probable cause, and is far from a prima facie case, let alone inculpatory evidence. Its intelligence value cannot exceed that given to police information or intelligence information and would therefore have no probative value. To use it as the sole source would violate the principle of presumption of innocence.

AI use should be directed towards the principle of beneficence or doing good, the betterment and progress of humanity. Thus, the development and use of AI systems must be directed to the benefit and welfare of society and human civilisation for the improvement of living conditions, health, work, development of physical and mental capacities.

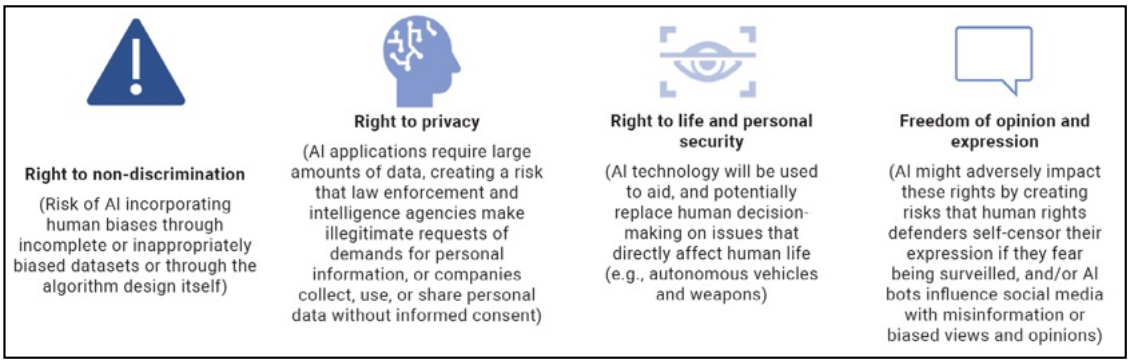
Although the basic structure and institutional framework for human rights protection, which is well-established and universally recognized, can be expected to develop effective responses to many of the threats and challenges wrought by the rising power of digital automation and machine intelligence, there are several reasons why the existing human rights enforcement mechanisms may require reinvigoration if they are to provide effective protection: First, many of the rights are difficult to assert in practice, due to the opacity of many of the socio-technical systems in which these technologies are embedded. Second, our understanding of the scope and content of existing rights was developed in a pre-networked age. So conceived, these rights might fail to provide comprehensive protection against the full range of threats and risks to individuals these technologies may give rise to, particularly in relation to discrimination and illegitimate attempts to deceive and manipulate individuals using "persuasive technologies"[251].

250 Ibid.

251 Persuasive technology is a "technology created specifically to change its users' opinions, attitudes, or behaviors to meet its goals", see: Centre for Humane technology (2021). Persuasive Technology. How does technology use design to influence my behavior?, available at: https://assets.website-files.com/5f0e1294f002b15080e1f2ff/612f8e3e010ff2e211c92019_2%20-%20Persuasive%20Technology%20Issue%20Guide.pdf

Figure 12 below gives an overview of some human rights covered in this toolkit that might be impacted by AI deployment in general.

Figure 12. Select human rights impacted by AI



Source: OECD, https://www.oecd-ilibrary.org/sites/ba682899-en/images/images/03_ba682899/media/image2.png

Interconnectivity of rights

Privacy is not just about access to data or collection thereof. Privacy is intimately linked to freedom of expression and association (because data reveal one's opinions and actions), freedom of movement (data points inform companies of one's location across time), and security (as data might expose individuals to repression and arbitrary detention).

The ACLU video below mentions how the DHS can use AI to determine whether an immigration applicant would become "a positive contributing member of society". This means that not only will AI be used to score immigration applicants, but it could further be used to score nationals (citizens and legal residents) once the technology is perfected.

In other words, AI could not just serve for surveillance purposes, but to police individual behaviors.

Activity: Training participants watch the video and discuss how AI can impact human rights.

Source: https://youtu.be/TbBMeFrR7H8

Advantages of human rights approach to AI development and deployment

Human rights law institutional mechanisms provide the direction and basis to ensure the ethical and human-centred development and use of AI in society. Judicial operators can recommend human rights due diligence such as human rights impact assessments (HRIAs) to assess and evaluate the risks posed by deployment of AI on human rights. The higher the risk to human rights, the more AI could be deemed unfit for use.

Human rights impact assessments can help identify vulnerable or at-risk groups or communities in relation to AI. Some individuals or communities may be under-represented due, for example, to limited smartphone use and the absence of their data in the datasets used to train AI systems. Human rights-based approach can provide remedy to those whose rights are violated. Examples of remedies include cessation of activity, development of new processes or policies, an apology or monetary compensation.

There are five key advantages in leveraging human rights frameworks in the AI context.

- Over time, a vast international, regional, and national human rights infrastructure has been developed, and there are established institutions that can help the realisation of human rights in the context of artificial intelligence. This infrastructure includes intergovernmental organisations, courts, NGOs, academic institutions, and other institutions and communities where human rights can be asserted, and redress sought.
- A comprehensive body of national, regional, and international law has operationalised the application of human rights in the digital realm.
- Human rights give a universal language for issues that transcend national borders, such as AI. Together with the human rights infrastructure, this can help to reach and include a broader range of stakeholders.
- Human rights enjoy widespread worldwide legitimacy and support. The mere perception that an actor may violate human rights might be significant due to the substantial reputational costs associated with such a perception.
- Many states have some form of a human rights framework, even if they do not have a data protection framework - therefore using the human rights framework as a base would make the process more inclusive.

A challenge related to human rights approach to AI development and deployment is the fact that their enforcement is tied to jurisdictions. Claimants must often demonstrate legal standing in a particular jurisdiction. When issues involve major international corporations and AI systems that span numerous jurisdictions, these approaches may not be optimal.[252]

252 Ibid.

Table 5. Key international instruments pertaining to the right to privacy in general, and in the online environment in particular.

Treaties	
1	Universal Declaration of Human Rights [253]
2	International Covenant on Civil and Political Rights [254]
3	International Convention on the Elimination of All Forms of Racial Discrimination [255]
4	OECD Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [256]
5	Council of Europe Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108 / Convention 108+) [257]
Standards	
6	The United Nations Guidelines concerning Computerized Personal Data Files (UN, 1990) [258]
7	The International Standards on Privacy and Data Protection (the Madrid Resolution) [259]
8	The OECD Recommendation on Digital Security Risk Management for Economic and Social Prosperity [260]
9	The OECD Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [261]
10	The UN Principles on Personal Data Protection and Privacy (2018) [262]
11	The UN General Assembly Resolution on the Right to Privacy in the Digital Age of 2014 [263]
Other documents	
12	The 2014 Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism [264]
13	The 2018 UN Promotion and protection of the right to freedom of opinion and expression [265]
14	The UN General Assembly Resolution on the right to privacy in the digital age (2020) has referred to "hacking and the unlawful use of biometric technologies," as "highly intrusive acts that violate the right to privacy" that interfere with freedom of expression and opinion, peaceful assembly and association, and the freedom of religion or belief, and "may contradict the tenets of a democratic society, including when undertaken extraterritorially or on a mass scale." [266]
15	A 2021 United Nations High Commissioner for Human Rights Report "The right to privacy in the digital age" has called for a moratorium on the deployment of facial recognition technologies in public spaces, until governments can show that there are no substantial issues related to the accuracy or discriminatory impacts and that these technologies comply with robust privacy and data protection standards. [267]
16	UNSDG Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda (2017) [268]
17	UN Compendium of data protection and privacy policies [269]

253 See: https://www.un.org/en/about-us/universal-declaration-of-human-rights
254 UN Human Rights Office (1976). International Covenant on Civil and Political Rights, available at: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights
255 UN Human Rights Office (1965). International Convention on the Elimination of All Forms of Racial Discrimination, available at: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial
256 OECD (2002). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, available at: https://www.oecd.org/en/publications/oecd-guidelines-on-the-protection-of-privacy-and-transborder-flows-of-personal-data_9789264196391-en.html
257 Council of Europe (CoE) (2018). Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108+), available at: https://www.coe.int/en/web/data-protection/convention108-and-protocol
258 Joinet L. (1988). Guidelines for the Regulation of Computerized Personal Data Files: final report, available at: https://digitallibrary.un.org/record/43365?ln=en&v=pdf
259 See: https://globalprivacyassembly.org/wp-content/uploads/2015/02/The-Madrid-Resolution.pdf
260 OECD (2015). Digital Security Risk Management for Economic and Social Prosperity: OECD Recommendation and Companion Document, available at: https://www.oecd.org/publications/digital-security-risk-management-for-economic-and-social-prosperity-9789264245471-en.htm
261 See: https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm
262 See: https://unsceb.org/privacy-principles
263 UN Human Rights Council (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, available at: https://digitallibrary.un.org/record/777869?v=pdf
264 See: https://www.ohchr.org/en/special-procedures/sr-terrorism
265 UN General Assembly (2018). Promotion and protection of the right to freedom of opinion and expression, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement
266 UN General Assembly (2020). The right to privacy in the digital age, available at: https://digitallibrary.un.org/record/3896430?v=pdf
267 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/calls-for-input/2021/right-privacy-digital-age-report-2021
268 See: https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda
269 UNDP (2021). COMPENDIUM OF DATA PROTECTION AND PRIVACY POLICIES AND OTHER RELATED GUIDANCE WITHIN THE UNITED NATIONS ORGANIZATION AND OTHER SELECTED BODIES OF THE INTERNATIONAL COMMUNIT, available at: https://unstats.un.org/legal-identity-agenda/documents/Paper/data_protecton_%20and_privacy.pdf

Right to access to court, fair trial, and due process

"All persons shall be equal before the courts and tribunals. In the determination of any criminal charge against him, or of his rights and obligations in a suit at law, everyone shall be entitled to a fair and public hearing by a competent, independent and impartial tribunal established by law [...] Everyone charged with a criminal offense shall have the right to be presumed innocent until proven guilty according to law."

– Article 14 of the ICCPR

When it comes to law enforcement and the legal system, the potential for AI to reinforce or amplify existing biases is a major concern. The rights to liberty, security, and fair trial may be infringed upon when an individual's physical freedom or personal safety is at stake, such as with predictive policing, recidivism risk assessment, and sentencing. As already discussed, "black box" AI systems make it impossible for legal professionals such as judges, attorneys, and prosecutors to comprehend the rationale behind the outcomes of the system, which complicates the justification and appeal of the decision.[270]

AI and Automated Decision Making (ADM) have a substantial impact on people's lives, and they might frequently restrict one's right to participate in, contest, or otherwise challenge the decision's outcome or its inputs. Often, AI systems, due to their "black box" nature, are unable to produce a human-intelligible and understandable explanation of their decisions. These systems can also have embedded biases that limit data invisible and marginalized groups' access to courts and justice.

Tools for criminal risk assessment, for instance, are offered as instruments to assist judges in sentencing decisions. Although authorities attribute a level of potential guilt by categorizing a person as high or low risk of reoffending, this could be at odds with the right to an impartial jury and the presumption of innocence. Predictive

270 CAHAI Secretariat (2020). Towards Regulation of AI Systems. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law, Council of Europe Study DGI/2020/16, available at: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.

policing software also may mirror societal biases and may pose the risk of using historical data to introduce bias and falsely attribute guilt.[271] There are several documented instances where the use of AI algorithms in predictive policing, risk assessment and sentencing has led to sub-optimal outcomes in the criminal justice system. In many cases, the use of AI for risk scoring of defendants and predictive policing efforts are advertised as well-intentioned attempts to remove the potential human bias of judges in their sentencing and bail decisions while allocating limited police resources to prevent crime. However, these AI systems, if not designed with ethical concerns in mind, may end up exacerbating the very bias they seek to mitigate by either directly incorporating biased factors or using proxies for bias in their recommendations.[272] This can result in serious consequences, including perpetuating discrimination against certain groups.

Therefore, when AI systems are biased, and opaque they raise concerns regarding fair trial standards, such as the presumption of innocence, the right to be informed promptly of the origin and nature of an accusation, the right to a fair hearing, and the ability to defend oneself in person. The opaqueness of decision making by AI systems also raises concerns regarding the arbitrary deprivation of liberty, and the right not to be punished without law.[273]

> *"Using risk assessment tools to make fair decisions about human liberty would require solving deep ethical, technical, and statistical challenges, including ensuring that the tools are designed and built to mitigate bias at both the model and data layers, and that proper protocols are in place to promote transparency and accountability. The tools currently available and under consideration for widespread use suffer from several of these failures".*[274]

271 AccessNow (2018). AI and human rights, available at: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

272 For instance, according to public records, the police in New Orleans used software created by Palantir for criminal investigations in a manner that extended beyond the software's original intended scope. Following a sequence of investigative reports and significant public backlash, the city terminated its six-year contract with Palantir in March 2018.

273 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, available at: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da.

274 Partnership on AI, Report on Algorithmic Risk Assessment Tools in the US Criminal Justice System, available at: https://partnershiponai.org/paper/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/

Activity: Training participants read select case law that deals with algorithmic black boxes in ADMs and AI systems and discuss how AI and technological advances affect human rights to access to court, fair trial and due process.[275]

State v. Loomis in the United States

In State v. Loomis, the Wisconsin Supreme Court determined that the use of the COMPAS algorithm, a proprietary risk assessment tool, at sentencing did not violate the defendant's due process rights. COMPAS was initially developed to assist parole boards in determining the risk of recidivism. However, the result of COMPAS—a risk assessment score—was used by both the State and the trial court during sentencing. Northpointe, Inc., the company that created COMPAS, refused to reveal its methodology to the court or the prisoner. The sentencing court gave the defendant a six-year sentence instead of parole since the algorithm determined that he had a significant probability of recidivism.[276]

Although the Court upheld COMPAS's validity, there were many limitations placed on its application. The algorithm could not be used to assess whether a criminal would serve time in prison or to estimate how long their sentence would be. Any Pre-sentence Investigation Reports including the score had to include an elaborate, five-part disclaimer about the algorithm's limitations. Its usage also required a separate justification for the sentence. The Supreme Court declined to take the case on appeal from the defendant.[277]

It remains an open question as to whether it is appropriate that the court permitted an algorithm, into which judicial operators have limited visibility, to play even a minor role in depriving a person of their liberty. The ruling of the Wisconsin Supreme Court and the appellate papers reveal fundamental errors regarding the potential operation of an algorithm like COMPAS and the protections necessary to make it useful in sentencing. These misunderstandings offer a glimpse into a more promising framework, one that would let algorithms strengthen the justice system without posing legal, technological, or ethical issues.[278]

People v. Alvin Davis in the United States

In this case, two witnesses claimed to have seen a black man in his mid-fifties on the property the day before the murder of an older woman who had been sexually attacked and murdered there. In the short months prior to the murder, dozens of persons, including Mr. Davis and another person, had visited the victim's residence. Mr. Davis is an African American man who had Parkinson's disease and was in his 70s at the time of the murder. A second person who fit the description of the witnesses had a history of sex offences.

Numerous sites and objects at the crime scene were sampled for DNA. Many of those items, including a cane that was allegedly used to sexually assault the victim, did not contain Mr. Davis' DNA. Although STRMix, a software used for DNA analysis, was able to successfully match Mr. Davis to the DNA sample from a shoelace that was probably used to tie up the victim, traditional DNA software was unable to do so. The prosecution extensively emphasized STRMix before the jury. Due to Parkinson's disease, Mr. Davis is confined to a wheelchair. The first trial against him ended in a hung jury. After a second trial, he was found guilty and given a life without parole term.

In People v. Alvin Davis in California, the Electronic Frontier Foundation (EFF) intervened in favour of Mr. Davis's ability to view the source code of STRMix, the forensic DNA programme that was employed during his trial. The EFF has claimed that a defendant has the right to review DNA analysis software in several cases, the most recent of which is this one. In two of those instances, United States v. Ellis and State v. Pickett, the courts agreed with EFF that the defendants were entitled to the TrueAllele source

275 Grimm P., Grossman M. R., Cormack G. V, Artificial Intelligence as Evidence, Artificial Intelligence as Evidence, 19 Nw. J. Tech. & Intell. Prop. 9, available at: https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss1/2

276 Israni E. (2017). Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis, available at: https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1

277 Ibid.

278 Ibid.

code, one of STRMix's key rivals.[279]

To ensure that the outcome of DNA matching software used against them is accurate, criminal defendants must be permitted to review how the software functions. Since there may be coding flaws, having access to the source code cannot be a substitute for testimony on how the software should operate. This is particularly true for the most recent forensic DNA software, such STRMix and TrueAllele, which is rife with issues of accuracy and trustworthiness.[280] In reality, STRMix was previously examined, and it was found to contain programming faults that may have resulted in erroneous results in 60 cases in Queensland, Australia.[281]

State of New Jersey v. Pickett; United States v. Ellis

In both New Jersey v. Pickett[282] and United States v. Ellis[283], the defence sought access to the software source code of a firm (TrueAllele). TrueAllele is used to do a probabilistic genotyping study on DNA samples. The courts in both cases concluded that access to the code should be granted to the defence contingent on a protection order. The court in Pickett emphasized, "anything less than full access contravenes fundamental principles of fairness, which indubitably compromises a defendant's right to present a complete defense." While these tools are different from data-driven AI technologies, rulings showing that software source code can be accessed in criminal proceedings set an encouraging precedent for other advanced technologies claiming trade secret protections.[284]

State of New Jersey v. Francisco Arteaga in the United States

New Jersey v. Arteaga is an example of a case that highlights the importance of discoverability of AI algorithms and their data inputs in court cases. In 2019, a business in West New York, New Jersey was robbed at gunpoint, and Francisco Arteaga was subsequently identified as the suspect and charged with the robbery. Prior to the identification of Mr. Arteaga, New Jersey police discovered that witnesses at the crime site were unable to identify the offender, and a face recognition search conducted by New Jersey's Regional Operations Intelligence Center, yielded no results.

After this unsuccessful attempt at identification of the suspects, the New York Police Department conducted a facial recognition search using still photos cut from street-level surveillance cameras. Mr. Arteaga was among the search results, and the NYPD's facial recognition analyst identified him as the "possible match". The police subsequently placed Mr. Arteaga's photo in a photo line-up, where two witnesses eventually identified him, notwithstanding the flawed processes used to conduct the line-ups. Despite the significance of the algorithm based matching to the case, the defence was not given any information regarding the algorithm that generated it. Mr. Arteaga demanded discovery on the facial recognition technology used by the NYPD, the original photo and any edits made by the NYPD prior to running a search, and information regarding the analyst who ran the search that identified him. The district court in New Jersey refused his plea to order discovery.

EPIC together with the Electronic Frontier Foundation (EFF) and the National Association of Criminal Defense Lawyers (NACDL) filed a brief informing the court about how errors occur in facial recognition systems, the potential for bias in those systems. They argued that discovery is the final opportunity to correct these errors. The brief outlined the sequence of procedures necessary to conduct a facial recognition search, all of which entail human decisions that can add inaccuracies and increase the likelihood of misidentification. The brief contends that human review following a search cannot be considered a remedy for algorithmic errors.[285] The case is now before the Judge of the Court of Appeal.

279 Zhao H. (2021). EFF tells California Court that Forensic Software Source Code Must Be Disclosed to the Defendant, available at: https://www.eff.org/deeplinks/2021/05/eff-tells-california-court-forensic-software-source-code-must-be-disclosed

280 Zhao H. (2021). How Your DNA—or Someone Else's—Can Send You to Jail, available at: https://www.eff.org/deeplinks/2021/05/how-your-dna-or-someone-elses-can-send-you-jail.

281 Murray D. (2015). Queensland authorities confirm 'miscode' affects DNA evidence in criminal cases, available at: https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1

282 State of New Jersey v Corey Pickett, available at: https://law.justia.com/cases/new-jersey/appellate-division-published/2021/a4207-19.html

283 EFF, United States v Ellis, available at: https://www.eff.org/cases/united-states-v-ellis

284 NACDL'S task force on predictive policing (2021). Garbage in, gospel out. How Data-Driven Policing Technologies Entrench Historic Racism and 'Tech-wash' Bias in the Criminal Legal System, available at: https://www.nacdl.org/Document/GarbageInGospelOutDataDrivenPolicingTechnologies

285 EPIC Amicus Brief, New Jersey v. Arteaga, available at: https://epic.org/documents/new-jersey-v-arteaga/

One of the greatest threats generated using AI systems in the administration of justice is the so-called automation bias, which is the tendency of humans to uncritically consider the solution offered by artificial intelligence as correct, producing an automatic validation by humans. This is a particularly aberrant risk in the administration of justice, as it can lead to blind trust in the system's proposed decisions, to consider the only existing jurisprudence to be that proposed by the machine, or to consider an assessment of the possibility of recidivism to be correct. Over time, this would lead to a change in the reasoning of decisions aimed at justifying why the result offered by the system is not followed, a possibility that is aggravated by the disproportionate workload of most of our courts, which leads to a system of work in which quantity and speed take precedence over quality. It is for this reason that the judge's departure from any decision, whether assisted or automated, cannot lead to any kind of reprisal, sanction, inspection or disciplinary regime. If human supervision and control prevails, the control must be effective. Key questions that we should ask in this regard:

(i) How does the resolution of a case by an AI system, instead of a judge affect the right to effective right to access to court, fair trial, and due process?

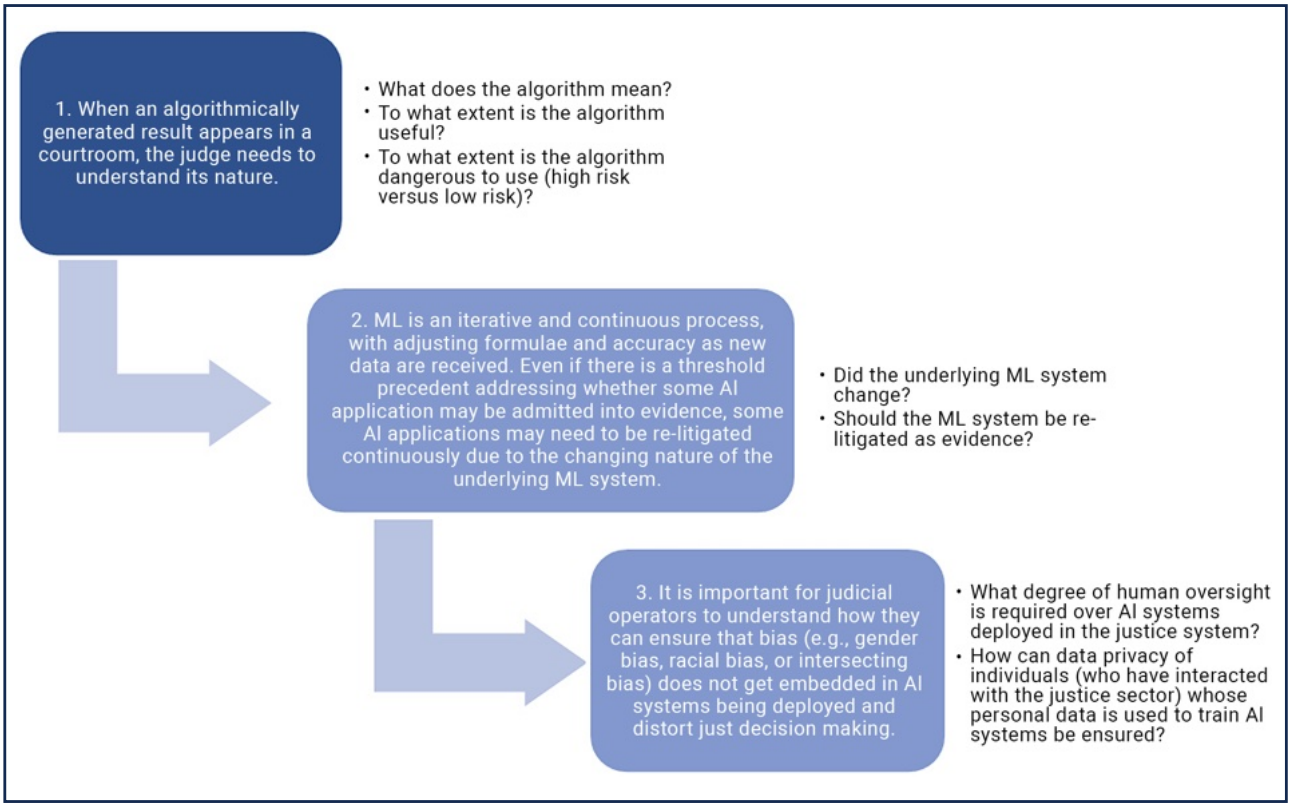
(ii) How will the motivation of judicial decisions be articulated? Citizens have the right to know the motivation of judgments and judges have the duty to give reasons for them. In the case of a black box, the logical reasoning of the conclusion is neither transparent nor can it be obtained.

(iii) In the case of the existence of a proposal for a draft decision or the application of case law by an AI system that feeds into a decision/judgement made ultimately by a human judge, do the parties have the right to know the reasoning of the AI system, and could that argument be used as a reason for appeal or as an argument to support the appeal? The right to transparency of the algorithm and the secret deliberations of the court are two separate issues that should not be confused as being the same.

AI systems should be considered as auxiliary and support tools, without attributing a decisive value to them or falling into overestimation, without forgetting the necessary judicial motivation and the essential individualization of sentences. The right not to be subject to a solely automated decision, the right to be informed of the automated decision, the right to challenge or review automated or algorithmic decisions, and the right to request human supervision and intervention should be guaranteed (see supra, p.78-82).

Figure 13 below outlines a few steps that judicial operators could follow when deciding cases that involve AI and human rights:

Figure 13. Steps that judicial operators could follow when deciding cases that involve AI and human rights



Source: Authors

Activity: Ensuring that AI systems are used in a manner that upholds the principles of a fair trial is crucial to maintaining the integrity of the legal system. Here is a hypothetical case example that illustrates the importance of AI in ensuring a fair trial. Please review the facts of the case and discuss what laws would have applied if the case was tried in your jurisdiction. What would have been the outcome of the case?

Case Title: The State vs. John Doe

Background: John Doe is facing criminal charges related to a robbery that occurred at a convenience store. The prosecution is relying on surveillance camera footage as a key piece of evidence. The defense, however, argues that the footage is inconclusive and that John Doe is being wrongfully accused.

Role of AI in Ensuring a Fair Trial:

1. Video Analysis AI: The prosecution introduces an AI-based video analysis system that claims to enhance and analyze the surveillance footage. This AI system is said to have the ability to identify faces, enhance image quality, and detect suspicious behavior.
2. Concerns Raised by the Defense: The defense raises concerns about the accuracy and potential biases of the AI system. They argue that the AI may have been trained on biased datasets and that its results might not be reliable.
3. Expert Witnesses: Both the prosecution and the defense call expert witnesses to testify about the AI system's capabilities and limitations. The defense's expert witness questions the AI's accuracy and highlights potential biases.
4. Transparency and Explainability: The defense requests that the AI system's algorithms and decision-making processes be disclosed for examination. They argue that without transparency and explainability, the AI's findings cannot be trusted.
5. Independent Review: The court orders an independent review of the AI system's output and algorithms by a neutral third party. This review aims to assess the accuracy and fairness of the AI's findings.
6. Legal Precedent: The case brings attention to the need for legal standards and guidelines regarding the use of AI in criminal trials. The court considers whether the use of AI in this case complies with existing legal standards and principles of fairness.

Outcome:

The court ultimately decides to admit the AI-enhanced video analysis as evidence, but with conditions:

- The AI system's algorithms and decision-making processes must be disclosed to the defense and the independent reviewer.
- The court acknowledges that AI systems can introduce bias and errors and that expert testimony regarding the AI's limitations will be allowed.
- The independent reviewer will assess the AI's findings and provide a report to the court.

This hypothetical case highlights the importance of transparency, fairness, and accountability when using AI in the legal system. It also underscores the need for legal standards and guidelines to ensure that AI technologies do not compromise the principles of a fair trial, including the right to a defense, the right to challenge evidence, and the right to examine and cross-examine witnesses.

Effective remedy

The use of AI systems in situations where human rights are at stake may present difficulties in ensuring the right to remedy. Since many AI systems are opaque, individuals may be unaware of how decisions affecting their rights were made, or whether the process was discriminatory. Often, the judicial operator using the AI system may be unable to explain the automated decision-making process (see supra, p.78-82). These issues are compounded by the deployment of AI systems that recommend, make, or enforce decisions within the Judiciary, the very institutions responsible for protecting rights, including the right to an effective remedy.[286]

Contestability

Affected individuals and groups should be afforded effective means to contest relevant determinations and decisions. As a necessary precondition, the existence, process, rationale, reasoning and possible outcome of algorithmic systems at individual and collective levels should be explained and clarified in a timely, impartial, easily-readable and accessible manner to individuals whose rights or legitimate interests may be affected, as well as to relevant public authorities. Contestation should include an opportunity to be heard, a thorough review of the decision and the possibility to obtain a non-automated decision. This right may not be waived, and should be affordable and easily enforceable before, during and after deployment, including through the provision of easily accessible contact points and hotlines.

Source: Council of Europe Recommendation CM/Rec (2020) of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers' Deputies)

Automated decision-making processes lend themselves to challenges for individuals' ability to obtain effective remedy. These include the opacity of the decision itself, its basis, and whether the individuals have consented to the use of their data in making this decision or are even aware of how it impacts them. It is unclear to whom persons should express their issues with the decision due to the difficulty in assigning responsibility for the decision. Due to the nature of judgments being made automatically, without or with little human input, and with a focus on efficiency rather than human-contextual reasoning, organizations deploying ADM systems have an even greater obligation to provide impacted individuals with a method to seek redress.[287] In this context, it is worth mentioning that the proposed EU AI liability directive would create a rebuttable 'presumption of causality', to ease the burden of proof to establish damage caused by an AI system. This will alleviate some of the hurdles when bringing a claim for harm caused by an AI system. It would furthermore give national courts the power to order disclosure of evidence about AI systems suspected of having caused damage.[288]

286 Toronto Declaration, available at: https://www.torontodeclaration.org/declaration-text/english/

287 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, available at: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

288 Proposal for a Directive on adapting non contractual civil liability rules to artificial intelligence, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0496

Activity: Training participants read select case law that deals with algorithmic black boxes in ADMs and AI systems and discuss how AI and technological advances affect the right to remedy.

People v. Chubbs (2015) in the United States

A California Appeals Court upheld a trade secret evidentiary privilege in a criminal case in 2015 to prevent the disclosure of TrueAllele source code to the defense. The ruling in People v. Chubbs is being referenced in the US to deny defendants access to trade secret evidence.[289] The court ruled that a defendant has no right to the source code of a DNA algorithm used to identify the defendant, prima facie. The owner of a trade secret has the right to refuse to disclose the secret if granting that right will not serve to hide fraud or otherwise promote injustice.[290] In this case, The California Court of Appeals extended a trade secret evidentiary privilege in a criminal case. It allowed the developer to "entirely" withhold the source code. The Chubbs case has formed the basis for a new body of case law in the US which denies access to the underlying source code of algorithms used throughout the criminal justice system.[291]

Uber case concerning the use of the fraud-detection programme "Mastermind" in Europe

One recent case against Uber has relied on Article 22 GDPR, which states that individuals "have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[292] The applicants requested that the Amsterdam District Court analyze Mastermind, Uber's sophisticated fraud-detection programme.

Invoking GDPR safeguards against automated decision-making, Uber drivers in the UK and Portugal claimed they were wrongfully fired by the company's anti-fraud algorithm. The applicants claimed that the algorithm used by Uber was automated (no meaningful human intervention) and resulted in termination of their job with Uber. Without giving them the possibility to challenge the decision taken by the company.[293]

The stated purpose of Mastermind is to assist Uber in effectively policing its platform. The lawsuit claimed that Uber has failed to demonstrate that its staff are knowledgeable enough about the inputs to its fraud fighting system to forecast the output or to explain the algorithm's judgments. It also stated that Uber is required to give drivers precise information about any alleged violations. According to the complaint, Uber's deactivation letters were mostly generic and omitted information concerning the alleged fraud. Additionally, the drivers were not given the chance to refute the allegations.[294]

289 Milner-Smith H., Copper D. (2017). When a computer program keeps you in jail, available at: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

290 People v. Superior Court of Los Angeles County (Chubbs) (Cal. Ct. App. 2015), available at: https://opencasebook.org/casebooks/2606-law-and-algorithms/resources/1.1-people-v-superior-court-of-los-angeles-county-chubbs-cal-ct-app-2015/

291 Chaney G. (2019). The Criminal Justice System's Algorithms Need Transparency, available at: https://www.law360.com/articles/1143086/the-criminal-justice-system-s-algorithms-need-transparency

292 https://ekker.legal/wp-content/uploads/2020/10/Court-request-Uber-account-deactivation-unofficial-translation.pdf. Article 22 of the GDPR: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

293 Huseinzade N. (2021). Algorithm Transparency: How to Eat the Cake and Have It Too, available at: https://www.europeanlawblog.eu/pub/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/release/1

294 Claburn T. (2020). Uber drivers take ride biz to European court over 'Kafkaesque' algorithmic firings by Mastermind code, available at: https://www.theregister.com/2020/10/26/uber_algorithmic_lawsuit/

A district court in Amsterdam has ordered Uber to reinstate drivers who were wrongfully terminated by the company's algorithm. It has also ordered Uber to compensate the drivers with more than €100,000 in damages.[295]

The case of Robodebt in Australia

In 2016, the Australian government introduced "Robodebt", an automated data-matching system to replace human examination of welfare recipients' income data. The goal was to detect overpayments or fraud. However, individuals flagged by the algorithm as suspicious were required to provide evidence to prove their innocence via an online form or risk losing their benefits entirely. This process has had detrimental effects on many individuals.

The algorithm, however, took tax authority data (which are based on a full year) and compared it with fortnightly income, ignoring the fact that the income of welfare recipients is often very irregular, due to short-term contracts or seasonal work. As a result, thousands of people were wrongly deprived of welfare payments, and many of them could not challenge these decisions as automated notifications were sent to an old address or they did not have access to the portal via which they could have forwarded the required evidence.

In many instances, people suddenly found themselves in serious debt, and even some cases of suicide were reported. Some sources calculate that the authorities have attempted to claim back almost 600 million AUD (360 million EUR) from citizens based on this system, which often generated errors but under which the burden of proof shifted to the individual. The results were very difficult to challenge. This case has reignited the debate on how algorithms and data matching are used to inform decisions.[296]

The proposed settlement for a class action against the Commonwealth of Australia regarding its use of Robodebt was approved by the Federal Court on June 11th, 2021. As per the settlement, the Commonwealth will pay $112 million (inclusive of legal costs) to certain group members as interest, refrain from raising, demanding or recovering any invalid debts from certain group members and accept court declarations that some of its administrative decisions were not validly made.[297]

295 Nawrat A. (2021). HR tech gone wrong? Uber told to reinstate drivers after 'robo-firing', available at: https://www.unleash.ai/hr-technology/court-rules-against-uber-robo-firing-employee-surveillance/

296 Human Rights Law Centre (2021). The Federal Court approves a $112 million settlement for the failures of the Robodebt system, available at: https://www.hrlc.org.au/human-rights-case-summaries/2021/9/30/the-federal-court-approves-a-112-million-settlement-for-the-failures-of-the-robodebt-system.

297 Ibid. See also: Katherine Prygodicz & Ors v The Commonwealth of Australia (No 2) [2021] FCA 634 (11 June 2021).

Rights to protection against discrimination

"All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status." – Article 26 of the ICCPR

"In those States in which ethnic, religious or linguistic minorities exist, persons belonging to such minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practise their own religion, or to use their own language." – Article 27 of the ICCPR

"The States Parties to the present Covenant undertake to ensure the equal right of men and women to the enjoyment of all [...] rights set forth in the present Covenant."

– Article 3 of the ICCPR and the ICESCR

The rights to protection against discrimination may be violated by AI systems, due to (i) the potential for bias on the part of algorithm developers; (ii) bias embedded in the model upon which the AI systems are built; (iii) bias embedded in the data sets used to train the models; or (iv) bias introduced when such systems are applied in real-world settings. These risks become exacerbated in situations when AI systems are deployed to assist judicial operators in their everyday activities.

The design of AI systems and their use in judicial procedures should be governed with the aim of producing human rights-compliant, non-discriminatory results. Minimum standards and safeguards should be established; if they cannot be met, the AI system in question should not be used.

Additionally, AI should be regulated so that it is sufficiently transparent and explicable to allow for effective independent review. The design and deployment of AI systems should comply with and give effect to the right to access the courts, the right to be presumed innocent, and the right to liberty, among others.

No human should be exposed to an automated decision that results in a criminal record, and AI technologies should not compromise the right to a fair trial by an impartial and independent tribunal. AI systems should not pre-label individuals as criminals without trial, nor should they enable the authorities to take unwarranted, disproportionate action against individuals without reasonable suspicion.

Where AI systems inform decisions on deprivations of liberty, they should be tuned to create outcomes that favour release, and they should not facilitate detention except as a last resort. To ensure that AI systems achieve the desired effect of lowering pre-trial detention rates, they must be subjected to rigorous testing.[298]

Issues that need to be considered by judicial operators when assessing the potential impact and risk of AI on the rights to protection against discrimination

- How, if at all, could the AI system result in discrimination, have discriminatory impacts on rights-holders, or perform differentially for different groups in discriminatory or harmful ways?
- How, if at all, could the use of the AI system exacerbate existing inequities or discrimination in the populations it affects?
- In what additional ways, if any, could the use of this system contribute to or exacerbate inequity or inequality?

Source: Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, available at: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf

AI systems must be designed to ensure that they do not produce discriminatory results, ensuring that suspects and accused individuals are not disadvantaged, either directly or indirectly, based on their characteristics, such as race, ethnicity, nationality, minority, or socioeconomic status. AI systems should be subject to mandatory testing before and after deployment to identify and correct any discriminatory effects. Please refer to Module 3 that discusses algorithmic bias in detail.[299]

AI systems must be transparent and comprehensible so that their key users, such as decision makers, parties to a litigation, defendants, can comprehend and scrutinise them. Commercial or proprietary interests, such as trade secrets, should be balanced with the requirements related to transparency.. Each AI system should be auditable by an independent auditor, and its processes should be replicable for this purpose.[300]

298 Fair Trials, Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU Policy Paper, available at: https://www.fairtrials.org/app/uploads/2022/01/Regulating-Artificial-Intelligence-for-Use-in-Criminal-Justice-Systems-Fair-Trials.pdf
299 Ibid.
300 Ibid.

Activity: Training participants read the facts of the Deliveroo and Foodhino cases and discuss how the opacity of AI algortihms and their functioning as black boxes affect rigths to protection against discrimination and pereptuate bias.

Deliveroo Case (2021)[301]

Deliveroo is a food delivery service that functions as three-sided marketplace through an online application. It connects local consumers, restaurants and grocers, and riders. Three labour unions challenged Deliveroo in Italian courts for violating regional labour laws. In this case, the Bologna court ruled that Deliveroo's reputational rating algorithm discriminated against food delivery workers or riders.[302] The court-examined ML algorithm was reportedly used to estimate a rider's "reliability." The court noted that the rider's "reliability index" would suffer if they failed to cancel a pre-booked shift using the app at least 24 hours before the start time. Since the algorithm prioritised offering shifts in busy time blocks to more dependable riders, riders who cannot fulfil their shifts, even in the event of a serious emergency or illness, will have fewer job options in the future. According to the court, the failure of the ML algorithm to consider the cause for a cancellation constituted discrimination and unfairly penalized riders who had legally valid reasons for not working. Deliveroo was ordered to compensate the plaintiffs with €50,000.[303]

The court also noted that the criteria for the algorithm's operation were neither defined on the app beyond generic aspects of reliability and participation, nor were they supplied to the court by the defendant corporation, which hindered a thorough assessment of the matter.[304]

Foodinho Case (2021)

Foodinho, another food delivery service based in Italy, was penalized 2.6 million euros by the Italian Data Protection Authority (Garante) for using discriminatory performance measurement algorithms in relation to its employees. The authority declared Foodinho in violation of the principles of transparency, security, and privacy by default and by design, and it held the company accountable for failing to take appropriate steps to protect the rights and freedoms of its employees (i.e., riders) from discriminatory ADM. In terms of algorithmic management of gig workers, the Garante's decision is a first of its kind. The Garante claimed that Foodinho's management had violated Article 22(3) of the GDPR.[305]

301 Colossa A. (2021). Algorithms, biases, and discrimination in their use: About recent judicial rulings on the subject, available at: https://www.ciat.org/ciatblog-algorithms-biases-and-discrimination-in-their-use-about-recent-judicial-rulings-on-the-subject/?lang=en

302 Lomas N. (2021). Italian court rules against 'discriminatory' Deliveroo rider-ranking algorithm, available at: https://techcrunch.com/2021/01/04/italian-court-rules-against-discriminatory-deliveroo-rider-ranking-algorithm/

303 Geiger G. (2021). Court Rules Deliveroo Used 'Discriminatory' Algorithm, available at: https://www.business-humanrights.org/en/latest-news/court-rules-deliveroo-used-discriminatory-algorithm/.

304 Ibid.

305 Milner-Smith et al. (2021). Italian Supervisory Authority Fines Foodinho Over Its Use of Performance Management Algorithms, available at: https://www.insideprivacy.com/gdpr/italian-supervisory-authority-fines-foodinho-over-its-use-of-performance-management-algorithms/.

In its ruling the Garante has stated that Foodinho engages in two different kinds of automated processing activities: one falls under the purview of the "excellence system," and the other is a component of the system that distributes orders based on an internal algorithm known as "Jarvis." The internal scoring method used by Foodinho to provide delivery slots to its riders is known as the "excellency system", which rates each rider. Drivers with higher ratings are given priority when determining delivery slots. In practice, this means that the "less excellent" drivers are excluded from the allocation of delivery slots if the "more excellent" drivers have already taken all the available delivery slots. The "excellence score" is determined by an automated statistical process that primarily considers customer and business partner feedback as well as delivery rates. Importantly, positive feedback is given less weight than negative feedback, and the system penalises drivers who fall short of the required delivery levels. The algorithm ("Jarvis") that assigns orders makes use of information including the riders' geographic whereabouts as determined by their GPS devices, the pick-up location, the delivery address, any special-order requirements, and the type of vehicle used. Jarvis assigns orders and fully automates the processing of this data. However, Foodinho did not specifically explain to the Garante how this algorithm is linked to the excellence system.[306]

Freedom of expression and access to information

"Everyone shall have the right to freedom of thought, conscience and religion. This right shall include freedom to have or to adopt a religion or belief of his choice, and freedom, either individually or in community with others and in public or private, to manifest his religion or belief in worship, observance, practice and teaching. No one shall be subject to coercion which would impair his freedom to have or to adopt a religion or belief of his choice."

– Article 18 of ICCPR and Article 18 of UDHR

"Everyone shall have the right to hold opinions without interference. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice."

– Article 19 of the ICCPR

Several international legal frameworks and guiding principles establish that the human rights to freedom of expression and access to information extend to the Internet. In 2011, the UN's Human Rights Committee issued General Comment No 34[307] stating that Article 19 of the ICCPR[308] protects

306 Ibid.
307 UN (2011). General Comment No 34, available at: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
308 UN (1976). International Covenant on Civil and Political Rights, available at: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

all forms of expression and the means of their dissemination, including all forms of electronic and internet-based modes of expression (including access to online information). This means that the principle of safeguarding the right to freedom of expression extends to the online space just as it does in the offline world.[309] In 2012, the UN Human Rights Council adopted a groundbreaking Resolution 20/8[310] to promote, protect, and ensure the enjoyment of human rights online. This resolution affirms the importance of upholding human rights in the digital age: "The same rights that people have offline must also be protected online, in particular freedom of expression, which is applicable regardless of frontiers and through any media of one's choice, in accordance with articles 19 of the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights.[311] Similarly, the 2018 UN Human Rights Council resolution on the promotion, protection and enjoyment of human rights on the Internet stated that "the same rights that people have offline must also be protected online, in particular freedom of expression"[312] and called upon all states to ensure these rights.

The special rapporteur's annual and thematic reports address various issues such as state surveillance of communications[313], safeguarding citizens' rights during elections[314], online hate speech[315], encryption and anonymity[316], children's right to express themselves[317], the role of the private sector[318] and digital access providers[319], the impact of artificial intelligence on citizens' rights[320], protecting journalists' freedom of expression[321], and preventing censorship while addressing online gender-based abuse[322].

309 UN (2011). General Comment No 34 (para 15), available at: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
310 UN Human Rights Council (2021). The promotion, protection and enjoyment of human rights on the Internet, available at: https://ap.ohchr.org/documents/dpage_e.aspx?si=a/hrc/res/20/8
311 UN (UNGA) (2012). The promotion, protection and enjoyment of human rights on the Internet, 16 July.2012, A/HRC/RES/20/8, available at: http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/RES/20/8
312 UN Human Rights Council (2018). The Promotion, Protection and Enjoyment of Human Rights on the Internet, available at: https://digitallibrary.un.org/record/1639840
313 UN Human Rights Council (2013). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, available at: https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/RegularSession/Session23/A.HRC.23.40_EN.pdf
314 UN Human Rights Council (2014). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion andexpression,availableat:https://documents-dds-ny.un.org/doc/UNDOC/GEN/G14/071/50/PDF/G1407150.pdf?OpenElement
315 UN General Assembly (2019). Promotion and protection of the right to freedom of opinion and expression, available at: https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/A_74_486.pdf
316 UN Human Rights Council (2015). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion andexpression,availableat:https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/095/85/PDF/G1509585.pdf?OpenElement
317 UN General Assembly (2014). Promotion and protection of the right to freedom of opinion and expression, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N14/512/72/PDF/N1451272.pdf?OpenElement
318 UN Human Rights Council (2016). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion andexpression,availableat:https://documents-dds-ny.un.org/doc/UNDOC/GEN/G16/095/12/PDF/G1609512.pdf?OpenElement
319 UN Human Rights Council (2017). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion andexpression,availableat:https://documents-dds-ny.un.org/doc/UNDOC/GEN/G17/077/46/PDF/G1707746.pdf?OpenElement
320 UN General Assembly (2018). Promotion and protection of the right to freedom of opinion and expression, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement
321 UN Human Rights Council (2012). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion andexpression,availableat:https://documents-dds-ny.un.org/doc/UNDOC/GEN/G12/137/87/PDF/G1213787.pdf?OpenElement7
322 See: http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21317&LangID=E; DigWatch (2023). Freedom of expression online in 2023, available at: https://dig.watch/topics/freedom-expression

Digital platforms are driven by algorithms that determine how to handle, prioritise, distribute, and delete or remove third-party information online. There is a possibility that these activities do not meet the legality, legitimacy, and proportionality standards for reasonable restrictions on freedom of expression. Moreover, breaches of personal information have a chilling effect on freedom of expression. People self-censor and alter their behaviour when they fear they are being observed or lack anonymity. This effect will be amplified by AI-powered surveillance, which can have a negative impact on free expression.

Activity: AI and Freedom of Expression

Participants watch the video and discuss the possible implications of AI on freedom of expression.

Source: UNESCO, https://www.youtube.com/watch?v=j0Oz54A68qo

In the digital world of today, the enjoyment of freedom of speech is governed in private, hybrid, and public areas shaped by private firms, government authorities, and users in varied, highly asymmetric power relationships. In addition, these digital ecosystems have set the path for new types of governance of expression, such as those moderated by AI systems on social media platforms to curate content on the user news feeds.

AI, content moderation and freedom of expression

Internet intermediaries moderate content on their platforms. This content moderation is often conducted outside of public view and is frequently carried out by opaque AI systems at scale, with no assurance of compliance with the international human rights framework. Such automated instruments may put restrictions on the right to freedom of expression and access to information, regardless of the technological method employed.[323] They may exclude from public discourse specific individuals, organisations, ideas, or forms of expression.

As the quantity of online information requiring moderation grows inevitably and exponentially, the main online platforms are investing heavily in AI systems to automate content moderation. A major impetus for this is content moderation laws being enacted worldwide that impose severe fines for noncompliance if online platforms fail to swiftly delete information that violates national intellectual property laws, as well as laws against hate speech and child pornography.[324]

One of the major problems associated with automating content moderation is that the AI technologies used for this are built on NLP technology that is domain specific, i.e., the technology will only identify the types of content on which it was trained. For example, an NLP system that has been trained to identify racist speech is incapable of identifying violent content. Moreover, even within a certain topic, NLP algorithms might not be able to comprehend detailed nuances of human speech, such as sarcasm and parody.[325] A system that can detect racist content in a blog article may not reliably recognize similar content in a tweet, resulting in a very high error rate for these technologies.[326]

> A resourceful guide in freedom of expression and access to information issues in the digital environment is UNESCO's **"Safeguarding freedom of expression and access to information: guidelines for a multistakeholder approach in the context of regulating digital platforms"**

To illustrate this point further, during the coronavirus outbreak, YouTube replaced many of its human content reviewers with AI algorithms charged with identifying and removing videos containing disinformation and hate speech. The content moderation experiment on the platform failed. AI algorithms censored users excessively, therefore tripling the

323 OSCE (2022). Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual, available at: https://www.osce.org/representative-on-freedom-of-media/510332

324 Raso F., Hilligoss H., Krishnamurthy V., Bavitz C., Kim L. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks, available at: https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights

325 Mindmatters (2021). Can the machine know you are just being sarcastic, available at: https://mindmatters.ai/2021/05/can-the-machine-know-you-are-just-being-sarcastic/.

326 Ibid. Also, see: Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, availableat:https://www.lawfareblog.com/assessing-impacts-ai-human-rights-its-not-solely-about-privacy-and-nondiscrimination.

rate of inaccurate content removals. YouTube rehired some of its human moderators after a few months.[327] Another example would be the case of content moderation specialist Kate Klonick who was banned from Twitter for publishing a tweet containing the phrase "I will murder you," which Twitter's algorithm deemed an encouragement to violence.[328] However, Klonick was not advocating violence in any way. She was only referencing a humorous exchange between Molly Jong-Fast and her husband, who was going to take her meal away.

It is worth noting that NLP tools are not yet as effective in languages other than English. As a result, automated tools may not be as accurate in evaluating non-English speakers, which can unfairly limit their freedom of expression. This is especially true for language translation tools, which can sometimes struggle with nuanced meanings and context. For instance, an incident occurred where an Israeli-Palestinian man was arrested after posting a picture on Facebook with the caption "good morning" in Arabic. However, Facebook's AI-powered translation tool inaccurately translated the caption to "attack them" in Hebrew or "hurt them" in English. Facebook later acknowledged the mistake and apologized to the man and his family for any inconvenience caused.[329]

Under investment in content moderation of non-English content by social media platforms, has long been covered by research[330] and press articles[331]. For years, Facebook's content moderation policy qualified the Arabic word "shaheed" (martyr) as inciting violence. When this word was identified by the algorithm, the content was automatically taken down without exception. However, the word "shaheed" has multiple meanings which do not glorify violence. This policy led to multiple false positives, and censoring of millions of accounts across the globe. On March 26, 2024, the Oversight Board recommended to end the systematic censorship of the word "shaheed" because of serious human rights violations for Arabic speaking people. The Board considered the ban "disproportionately restricts freedom of expression [and] is unnecessary"[332]. Ultimately, in July 2024, Meta agreed to comply with the recommendation[333].

327 Ibid. Also, see: Vincent J. (2020). YouTube brings back more human moderators after AI systems over-censor, available at: https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns

328 Kloncik K. (2020). What I Learned in Twitter Purgatory, available at: https://www.theatlantic.com/ideas/archive/2020/09/what-i-learned-twitter-purgatory/616144/; Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, available at: https://www.lawfaremedia.org/article/assessing-impacts-of-ai-on-human-rights-it-s-not-solely-about-privacy-and-nondiscrimination

329 Hu X., Neupane B., Flores Echaiz L., Sibal P., Rivera Lam M. (2019). UNESCO Report Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective, available at: https://unesdoc.unesco.org/ark:/48223/pf0000372132

330 Griffin R., (2021) Frances Haugen's leaks should show us the real significance of digital sovereignty, Sciences Po, Chair Digital, Governance, Sovereignty, available at: https://www.sciencespo.fr/public/chaire-numerique/en/2021/10/26/frances-haugens-leaks-should-show-us-the-real-significance-of-digital-sovereignty/

331 Brandom R., (2018) Activists from Myanmar and beyond call for Facebook to fix moderation, The Verge, available at:, https://www.theverge.com/2018/5/18/17369570/facebook-coalition-myanmar-syria-sri-lanka-moderation-transparency-audit ; Scheck J., Purnell N., Horwitz J., (2021) Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show, The Wall Street Journal, available at: https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953

332 Oversight Board, (2024) Oversight Board Publishes Policy Advisory Opinion on Referring to Designated Dangerous Individuals as "Shaheed", available at: https://www.oversightboard.com/news/oversight-board-publishes-policy-advisory-opinion-on-referring-to-designated-dangerous-individuals-as-shaheed/

333 Oversight Board, (2024) Meta Accepts Key Oversight Board Recommendations to End Blanket Ban on "Shaheed", available at: https://www.oversightboard.com/news/meta-accepts-key-oversight-board-recommendations-to-end-blanket-ban-on-shaheed/

Recent studies from the Center for Democracy & Technology (CDT) have examined the content moderation policies and practices of Maghrebi Arabic[334] and Kiswahili[335] (a language spoken by over 100 million people in East Africa). The research has shown that companies actively select the languages they want to focus on, leaving other dialects free from content moderation or subject to limited moderation. In the case of Maghrebi Arabic, content moderators sometimes failed to understand the nuances of various dialects, leading to errors and inaccuracies.

These studies also highlight one specific challenge regarding the automation of content moderation of underrepresented languages although they are used by millions of people. The scarcity of available data makes it difficult to train and improve algorithmic systems, which leads to poor content moderation decisions.[336]

On content moderation disparities:

Research has shown that social media platforms apply dual standard content moderation rules to people that negatively impact minorities.[337] Indeed, social media companies developed a wide range of sanctions to restrict access to content (label, deletion, shadow banning, demonetization) and user accounts (functionality restriction, temporary suspension, permanent ban). Marginalized groups see their content massively taken down while dominant groups benefit from clemency and flexible approaches (labels, demonetization). Thus, the protection of freedom of expression reveals deliberate choices to favor certain categories of individuals and viewpoints.

A Guardian article dated from August 2024 shows the disparity in algorithmic and human content moderation in Hebrew and Arabic.[338] As noted by Facebook's independent due diligence report published in 2022, different content moderation standards are found to be applied for content in different languages.[339]

Content moderation, freedom of expression and association

Social media platforms often serve as an aggregating factor between individuals. Indeed, online publications often prompt various reactions, from liking and sharing to joining a group, commenting on a political matter, and joining a protest[340]. Content moderation manipulation can affect

334 Elswah M., (2024) Moderating Maghrebi Arabic Content on Social Media, CDT, available at:, https://cdt.org/insights/moderating-maghrebi-arabic-content-on-social-media/

335 Elswah M., (2024) Moderating Kiswahili Content on Social Media, CDT, available at: https://cdt.org/insights/moderating-kiswahili-content-on-social-media/

336 See UNESCO's work with the tech industry to preserve linguistic diversity, Language Technologies for All – LT4All 2025, available at: https://www.unesco.org/en/articles/language-technologies-all-lt4all-2025 ; see also, GIZ, "Articial Intelligence for All – FAIR Forward", available at: https://www.giz.de/expertise/html/61982.html

337 Díaz A., Hecht-Felella L., (2021) Double Standards in Social Media Content Moderation, Brennan Center for Justice at NYU, available at:, https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

338 Paul K., (2024) Meta struggles with moderation in Hebrew, according to ex-employee and internal documents, The Guardian, available at: https://www.theguardian.com/technology/article/2024/aug/15/meta-content-moderation-hebrew

339 Sissons M., (2022) An Independent Due Diligence Exercise into Meta's Human Rights Impact in Israel and Palestine During the May 2021 Escalation, Meta, available at: https://about.fb.com/news/2022/09/human-rights-impact-meta-israel-palestine/

340 Ashraf C., (2020), Artificial Intelligence and the Rights to Assembly and Association, Journal of Cyber Policy, 5 (2): 163–79, doi:10.1080/23738871.2020.1778760

what individuals see and thus neutralise the idea of shared concerns and interests which is at the source of assembly and association.[341] In other words, content moderation offers an unprecedented capacity to tolerate or deter individuals from exercising their rights to free access to information, expression, association, as well as privacy.[342] Thus, the nexus between these rights is paradoxically both strong and fragile.

Individuals can refrain from participating in public protests or sharing information out of fear that pictures may be taken and further harm their career or job opportunities.

Indeed, some AI systems scan social media posts and pictures to identify people and collect geolocation data to predict potential trouble or threats of violence. Cases of individuals arrested for participating in peaceful protests through facial recognition systems are numerous.[343]

The Pulitzer Center's AI Accountability Network has revealed that a company promoted its AI system towards universities to "forestall potential volatile protests/demonstrations".[344] Some universities used this AI tool to monitor their students and control protests related to abortion or travel bans. In North Carolina, students who criticized their university were subjected to surveillance using an AI system.[345]

Activity: Training participants read the case Gonzalez versus Google and discuss what laws would be applicable in their jurisdictions under these circumstances. Would this impact the outcome of the case?

In 2023, the US Supreme Court was presented with an interesting case, Gonzalez v. Google. The case was brought up after the tragic death of 23-year-old Nohemi Gonzalez in the Paris terror attacks in 2015. The family of Nohemi Gonzalez sought to hold Google accountable for its role in the attacks under the Anti-Terrorism Act, which allows families of those killed by terrorists to pursue legal action against those who "aid and abet" such groups. Initially the Supreme Court declined to rule on the case, specifically on whether targeted recommendations by social media algorithms would be excluded from the protection of Section 230 of the Communications Decency Act. This decision has implications for the future of liability in similar cases.

Source: https://www.supremecourt.gov/opinions/22pdf/21-1333_6j7a.pdf

341 Art. 20 UDHR ; art. 21-22 ICCPR

342 UNESCO, (2019) Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective, available at: Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective - UNESCO Digital Library

343 See; International Center for Not-for-Profit Law (ICNL), The Impact of Artificial Intelligence Technologies on the Right to Privacy and Civic Freedoms, 28 May 2021, p.2-3 https://www.ohchr.org/sites/default/files/Documents/Issues/DigitalAge/Submissions/CSOs/ICNL.pdf

344 Gaumond E., Régis C., (2023) Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, Lawfare, available at: https://www.lawfaremedia.org/article/assessing-impacts-of-ai-on-human-rights-it-s-not-solely-about-privacy-and-nondiscrimination

345 Swan S., Sen A., Bennett D., (2022) Universities and Students Investigate Use of AI Surveillance Tool in Response to Grantee's Reporting, Pulitzer Center, available at: Universities and Students Investigate Use of AI Surveillance Tool in Response to Grantee's Reporting | Pulitzer Center

In an environment where social media platforms utilize algorithms to decide whose voices we hear, the right to freedom of expression is of particular importance. In 2014, Cornell University researchers conducted an emotional contagion study in collaboration with Facebook, studying how emotions spread over the social network.[346] The researchers modified the experiences of over 700,000 Facebook users by employing a sentiment analysis technique to determine whether friends had contributed unpleasant comments or posts. These negative items were subsequently removed from users' newsfeeds in an experiment to determine whether algorithmically skewing the feed towards positive content would keep users on the site for longer. This study highlights how platforms might make decisions based on user expressions that support one reality and diminish another.[347]

The capacity of social media platforms to effectively control the emotions of their users is challenged[348]. However, these experiments and decisions can be deemed deceptive and manipulative, and have implications on individual users as they influence their social, economic or political behaviour.[349] Additionally, these practices contribute to the spread of hate speech or disinformation, two categories known for incentivizing reactions online (reposts, comments, likes).

Activity: "The Napalm Girl" content moderation case

The "Napalm Girl" content moderation case refers to a controversial incident involving the moderation of historical and iconic war-related imagery on social media platforms. The case revolves around the removal or censorship of a Pulitzer Prize-winning photograph known as "The Terror of War," which depicts a young girl, Kim Phúc, fleeing from a napalm attack during the Vietnam War. Training participants read the overview of the case and discuss its implications for the freedom of expression in the digital environment.

Background:

- The photograph was taken by Associated Press (AP) photographer Nick Ut on June 8, 1972, during the Vietnam War. It captures the immediate aftermath of a napalm bombing in Trang Bang, South Vietnam.
- The image features a naked, severely burned nine-year-old girl, Kim Phúc, running down a road in agony.
- The photograph has become an iconic symbol of the horrors of war and has played a significant role in raising awareness about the Vietnam War's human cost.

Content Moderation Incident:

- In September 2016, Facebook temporarily removed the photograph when it was posted by Norwegian writer Tom Egeland as part of a series on iconic war photographs.

346 Hu X., Neupane B., Flores Echaiz L., Sibal P., Rivera Lam M. (2019). UNESCO Report Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective, available at: https://unesdoc.unesco.org/ark:/48223/pf0000372132.

347 Meyer R. (2014). Everything We Know About Facebook's Secret Mood-Manipulation Experiment, available at: https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

348 Parmentier J., (2021) "Digital platforms have poor control over their manipulation of emotions", Polytechnique Insights, available at: https://www.polytechnique-insights.com/en/braincamps/digital/of-machines-and-men-new-challenges-of-digital-technology/digital-platforms-have-poor-control-over-their-manipulation-of-emotions/

349 Fischer A., (2022), Manipulation and the Affective Realm of Social Media, in Fleur Jongepier, Michael Klenk, The Philosophy of Online Manipulation, Routledge, (pp.327-352), available at: https://library.oapen.org/bitstream/id/2bf04b4d-5c54-467d-b5eb-4db6148e7fc7/9781000603583.pdf

- Facebook's reason for removal was its policy against displaying nudity on the platform.
- The decision sparked outrage and controversy, with many arguing that the photograph's historical and journalistic significance should outweigh concerns about nudity.
- After significant public backlash and criticism, Facebook reversed its decision and reinstated the photograph.

Key Issues and Debates:

1. Freedom of Expression vs. Content Moderation: The case raises questions about the balance between freedom of expression, the sharing of historical and newsworthy content, and the need for content moderation to prevent the spread of inappropriate or offensive material.
2. Algorithmic Moderation: Many social media platforms use algorithms to automatically detect and remove content that violates their policies. In this case, the algorithms failed to distinguish between a historic, Pulitzer Prize-winning photograph and inappropriate content.
3. Cultural Sensitivity and Context: Critics argue that content moderation algorithms lack the ability to understand the historical, cultural, and contextual significance of certain images, leading to erroneous removals.
4. Responsibility of Tech Companies: The incident also brings into question the responsibility of tech companies to make nuanced decisions about content moderation and the potential impact of their decisions on free speech and historical documentation.

Ultimately, the "Napalm Girl" content moderation case highlights the challenges faced by social media platforms and tech companies in striking a balance between moderating content to uphold community standards and recognizing the importance of historical and journalistic content, especially when it depicts sensitive or distressing subjects. It underscores the need for thoughtful, context-aware content moderation policies and decisions.

Source: Content Moderation Case Study: Facebook Attracts International Attention When It Removes A Historic Vietnam War Photo Posted By The Editor-in-Chief Of Norway's Biggest Newspaper (2016), available at: https://www.techdirt.com/2020/11/20/content-moderation-case-study-facebook-attracts-international-attention-when-it-removes-historic-vietnam-war-photo-posted/

Misinformation and AI

As already noted, AI technologies can contribute to unequal access to information and exacerbate existing digital divides. For instance, AI may be used to develop and spread targeted propaganda, and this problem is exacerbated by AI-powered social media algorithms driven by "engagement" that promote information that is most likely to be clicked. The data analysis used by social media businesses to construct user profiles for targeted advertising is powered by ML algorithms. In addition, bots masquerading as genuine users propagate content outside of tightly targeted social media groups by distributing links to bogus sources and communicating actively with people as chatbots using natural language processing.[350]

350 Ibid.

Entities deploying AI screening and scoring algorithms frequently fail to offer proper notification, if any, to those being scored and screened. Because consumers are unaware of how these tools make determinations and what types of data they employ, their use can erode restrictions relating to access to information. Because individuals do not understand how these tools function, they are unable to challenge eligibility decisions affecting their access to services, jobs, housing, or benefits.[351]

Moreover, the threat of deepfakes, which are AI systems capable of making realistic-sounding video and audio recordings of actual people, has led many to believe that the technology will be used in the future to make fake footage of world leaders for harmful purposes. Although it appears that deepfakes have not yet been used as part of actual propaganda or disinformation campaigns, and the forged audio and video are not yet convincingly human, the AI behind deepfakes is advancing, and the potential for spreading chaos, inciting conflict, and furthering the crisis of truth should not be discounted.[352]

In nations where religious liberty is threatened, AI could aid government officials in monitoring and targeting members of persecuted religious organizations. Not only may this increase the secrecy of such gatherings out of fear of being detected, but it could also result in physical consequences ranging from arrest to death. Additionally, AI might be used to identify and remove religious content. If people are unable to show religious symbols, pray, or teach about their faith online, this would be a flagrant infringement of the freedom of religion.[353]

The NGO AccessNow points out that online harassment enabled by bots poses a clear and imminent threat to free speech. These bot accounts pose as human users and deliver automatic responses to designated accounts or anyone who shares a particular viewpoint. This type of unrelenting online harassment has a chilling impact on free speech, especially for underprivileged groups that are disproportionately targeted. Bot developers apply natural language processing more frequently, which exacerbates online harassment threats by bots. This will make it more difficult to identify, report, and eliminate bot accounts.[354]

Legitimate restrictions to freedom of expression and access to information

In the international human rights framework and in numerous constitutions, there are stringent conditions for justifying prior limits on freedom of expression and access to information. In this aspect, AI tools are especially worrisome because these systems are hidden from public scrutiny, are

351 See: https://epic.org/issues/ai/screening-scoring/
352 AccessNow (2018). AI and human rights, available at: https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdf
353 Ibid.
354 Ibid.

context-blind, and function in a highly opaque manner that precludes any effective correction or retribution. While pre-screening content to restrict the online transmission of malware and child sexual abuse has been widely regarded as a useful application of automation, caution must be exercised when applying the same rationale to other sorts of speech that belong under the broader category of content regulation.[355] International law allows for the restriction of digital rights (the right to privacy, freedom of expression and access to information) both offline and online but only under very limited and specific circumstances, and in accordance with Article 19 of the ICCPR (freedom of expression and access to information) using the three-part test outlined below.[356]

Table 6. Three-part test for legitimate limits to freedom of expression

Principle	Explanation
The restrictions should be provided by law	• ICT laws must clearly stipulate any restrictions on freedom of expression unambiguously. Citizens must be able to understand and comply with the laws, making them legitimate. Vague and overly broad provisions would not meet this standard. • The Human Rights Committee of the United Nations has stated in General Comment No. 34 that restrictions on digital rights should be specific to the content. Broad bans on certain sites and systems are not in line with international law. Additionally, prohibiting the publication of material solely based on its criticism of the government or its political and social system goes against international law.[357]
The restriction should pursue a legitimate aim	• According to Article 3)19) of the ICCPR, limitations should only be imposed for legitimate reasons such as protecting the rights and reputations of others, ensuring national security, maintaining public order, and promoting public health or morals.
The restriction should be necessary for a legitimate purpose	• Any limitations to the right to freedom of expression must be necessary and proportionate. While public surveillance may be permissible, States must demonstrate that measures are both necessary and proportionate. Digital surveillance is a very intrusive act that violates digital rights. Prior approval from a competent judicial authority is necessary for proportionate digital surveillance. This also means that the least intrusive surveillance methods shall be used.[358] • For instance, automated threat detection, a system commonly utilized by police forces to detect gunshots and identify possible crime scenes, has been found to inaccurately identify sounds as gunshots in 89% of cases. Many police departments that previously utilized predictive policing services have discontinued these systems due to their limited usefulness and accuracy.[359]

355 Ibid.
356 See: UNESCO (2021). Global Toolkit for Judicial Actors: International legal standards on freedom of expression, access to information and safety of journalists, Module 2, 44–46, available at: https://unesdoc.unesco.org/ark:/48223/pf0000378755
357 UN (2011). General Comment No 34, available at: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
358 International Commission of Jurists, Regulation of Communications Surveillance and Access to Internet in Selected African States, disponible en: https://www.kas.de/documents/275350/0/Report-on-Regulation-of-Communications-Surveillance-and-Access-to-Internet-in-Selected-African-States.pdf/66dbd47d-4d7d-2779-a595-a34e9f93cfbb?t=1639140695434
359 Ibid.

The following video by UNESCO explains the three-part test for legitimate limits to freedom of expression:

Source: https://www.youtube.com/watch?v=Wg8fVtHPDag

Right to privacy and data protection

"No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."

- Article 17 of the ICCPR

Privacy is instrumental in securing other human rights including the rights to freedom of speech, opinion, affiliation, and assembly. Without privacy, it is often not practical or safe to organize political opposition, compete commercially or otherwise develop alternatives to existing policies, dominant narratives, or experienced injustice. The Universal Declaration of Human Rights (UDHR, article 12), the International Covenant on Civil and Political Rights (ICCPR, article 17), and several other international and regional human rights treaties acknowledge the right to privacy as a human right (see, supra, Table 5, p.135).[360] The significance of the right to privacy for the online and offline exercise of other human rights, such as the freedom of expression and access to information, is increasing in a data-centric world.[361]

360 For example, the Convention on the Rights of the Child (article 16), International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (article 14), Convention on the Rights of Persons with Disabilities (article 22), African Charter on the Rights and Welfare of the Child (article 10), American Convention on Human Rights (article 11) and Convention for the Protection of Human Rights and Fundamental Freedoms (the European Convention on Human Rights, article 8).

361 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/calls-for-input/2021/right-privacy-digital-age-report-2021

Since the ICCPR entered into force in 1976, new digital technologies have evolved, and governments and private organizations have more often than not exploited them outside of the legal framework and with disregard for individual privacy. While digital surveillance and digital technologies have advanced swiftly, the law on privacy has not followed suit. Although privacy legislation at the international human rights level is based on robust and well-established principles, it has not been evolved or modified to meet the requirements of 21st-century society. The UN Human Rights Committee's original 1988 General Comment on privacy did not anticipate the development of new forms of communication such as email and texting, the emergence of government capabilities to intercept and process large quantities of electronic data, or the explosion of social media websites, to name a few examples.[362]

The UN General Assembly Resolution on the right to privacy in the digital age (2020) has referred to "hacking and the unlawful use of biometric technologies," as "highly intrusive acts that violate the right to privacy" that interfere with freedom of expression and opinion, peaceful assembly and association, and the freedom of religion or belief, and "may contradict the tenets of a democratic society, including when undertaken extraterritorially or on a mass scale."[363] A 2021 UN High Commissioner for Human Rights Report "The right to privacy in the digital age" has called for a moratorium on using facial recognition technologies in public spaces, until governments can show that there are no substantial issues related to the accuracy or discriminatory impacts and that these technologies comply with robust privacy and data protection standards.[364]

Privacy and data protection in the digital realm

Understanding data protection and privacy law in the digital realm requires a comprehensive understanding of privacy's definition, classification, and emergence as a social concern. The right to privacy is a core principle for a democratic society and plays a crucial role in the balance of power between government, private sector entities that collect, process and store personal data, and individuals whose personal data are collected, processed, and stored. The significance of the right to privacy for the online and offline exercise of other human rights, such as the freedom of expression and access to information, is increasing in a data-centric world.[365]

362 American Civil Liberties Union (2015). Information Privacy in the Digital Age, available at: https://www.aclu.org/other/human-right-privacy-digital-age

363 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/calls-for-input/2021/right-privacy-digital-age-report-2021

364 Ibid.

365 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/calls-for-input/2021/right-privacy-digital-age-report-2021

Since the onset of the information age, the right to privacy and the necessity to protect personal information or data have received considerable attention. We live in an era in which digital technologies enable cost-effective mass collection, storage, and processing of personal data online, as well as monitoring of individuals wherever they are located (including monitoring of their online activities). While the internet and online information-sharing and data collection increase at an exponential rate, legislative developments have failed to keep pace and adequately protect personal information. Governments around the world have begun to adopt data protection-related instruments and regulations to protect the privacy rights of their citizens.[366]

The concept of privacy is a constellation of principles. The right to privacy guarantees that a space is reserved for self-expression. In this manner, the right is strongly related to freedom of expression. There is increasing recognition that the right to privacy plays a vital role in facilitating the right to freedom of expression and access to information. For instance, protection of the right to privacy allows individuals to share views anonymously in circumstances where they may fear being censured for those views, it allows whistle-blowers to make protected disclosures, and it enables members of the media and activists to communicate securely beyond the reach of government surveillance.[367] Additionally, the right to privacy safeguards intimacy and dignity. Furthermore, it includes the right to decide how to live and the right to autonomy in general. The right to privacy includes informational privacy, as well as the right to access and control one's personal information, regardless of its format. These privacy subcomponents are not exhaustive; rather, they serve as a roadmap for the future development of privacy measures in the digital environment.[368]

The line between the online and offline world is becoming increasingly blurred. Making it harder to define clear boundaries.

Social media companies thrive on the collection and commercialization of large volumes of Internet user data, which further emphasizes the need to protect user privacy. AI facilitates the collection, processing, and reuse of massive quantities of data and images, encouraging organizations (both private sector and government) to gather, retain, and handle sensitive data about users. AI algorithms make predictions regarding people's personal life, including things like where they live and their sleeping habits.

366 Media Defence (2022). Module 4: Data Privacy and Data Protection, available at: https://www.mediadefence.org/ereader/publications/modules-on-litigating-freedom-of-expression-and-digital-rights-in-south-and-southeast-asia/module-4-data-privacy-and-data-protection/introduction/

367 Ibid.

368 American Civil Liberties Union (2015). Information Privacy in the Digital Age, available at: https://www.aclu.org/other/human-right-privacy-digital-age

As we go about our daily lives, our smartphones' GPS trackers can collect a wealth of data about our movements, even if we're not actively using the internet. When we visit places like coffee shops, schools, and medical facilities, this information can be used to make inferences about our personal identity, interests, aspirations, problems, and social networks based on how long we stay and the movements of others around us. This data can be quite revealing and can have significant implications for our privacy and security. For example, when we move around the city and go to a coffee shop, a school, or a medical institution, the GPS tracker on our smartphones is able to detect where we are and how long we stay and collect this data (and correlate it with the movements of others), even if we did not access the Internet on our phones. Meaningful inferences can be derived regarding our identity, interests, aspirations, problems and networks from such data.

New and inexpensive forms of data analytics and storage coupled with enhanced digital and online connectivity (from smart appliances to nanobots inside human bodies) and emerging technologies such as AI and the IoT have enabled governments and giant corporations to become data miners, collecting information about every aspect of human activities, behavior, and lifestyle.

Privacy regulations have adapted to the novel challenges posed by the digital and online environment. Many nations around the globe have implemented regulations requiring data subjects' consent to use and process their personal data online, ensuring access to personal data by data subjects, and giving the right to have this personal data deleted, corrected, or transferred to a different entity.

Privacy-preserving laws in the AI environment aim to equip individuals with the right to view the content of databases containing information about them, and to oppose access to their personal information without their consent, except under limited circumstances defined by law. Under these laws, individuals have the right to agree to the terms of use before they download an app onto their cell phone or begin to use freeware, i.e., products and services whose economic model rests on commercializing personal data.[369]

369 Altshuler T. S. (2019). Privacy in a digital world, available at: https://techcrunch.com/2019/09/26/privacy-queen-of-human-rights-in-a-digital-world

Personal data, which is stored online, is often processed in numerous ways and purposes, some of which cannot be anticipated at the time when consent is granted by the data subject. Furthermore, many of us rarely go through the terms of use, even when they are concise and displayed in large print.[370] For instance, it will take us 76 days to read the privacy policies that one may encounter every year.[371]

Another aspect of privacy in the AI environment is understanding privacy as the "right to be left alone."[372] This refers to the right to keep a safe and protected space around our body, intimate thoughts, feelings, and lifestyle when being online. Constant online monitoring of our actions by sensors, surveillance cameras, digital assistants, such as Siri, Alexa, and other AI and digital tools, can have a profound impact on the right to privacy as a human right.[373]

Case Study: Amazon Alexa Recording and Sending Private Conversations

A family in Oregon, USA, reported that their Amazon Echo device had recorded a private conversation they were having in their home. Even more concerning, the recorded conversation was then sent to a contact in the family's address book, a colleague of one of the family members, without their consent or knowledge. The incident came to light when the recipient of the recorded conversation contacted the family to inform them about the unusual message. Amazon investigated the incident and attributed it to an extremely rare combination of circumstances. According to Amazon, the Echo device had mistakenly interpreted parts of the conversation as commands to send a message. It was a case of "false positive" wake word detection, where the device mistakenly thought it heard the wake word (likely "Alexa") and began recording and sending the conversation. Amazon took the incident seriously and took steps to improve the wake word recognition technology to prevent such false positives. The company also introduced a feature that allows users to add a PIN to voice purchases to prevent accidental orders through voice commands. This incident prompted discussions about the privacy and security of voice-activated devices, leading to increased awareness and user concerns about the potential for eavesdropping. Overall, the incident highlighted the need for technology companies to continually enhance the privacy and security features of voice-activated devices like Amazon Alexa. It also emphasized the importance of user education regarding device settings and privacy controls to ensure a safer and more secure user experience.

Source: Wolfson S. (2018). Amazon's Alexa recorded private conversation and sent it to random contact, available at: https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation

370 Ibid.
371 Popkin H. A. S. (2012). Life is too short to read privacy policies - here's statistical proof!, available at: https://www.nbcnews.com/tech/tech-news/life-too-short-read-privacy-policies-heres-statistical-proof-flna297399
372 Altshuler T. S. (2019). Privacy in a digital world, available at: https://techcrunch.com/2019/09/26/privacy-queen-of-human-rights-in-a-digital-world
373 Ibid.

It is important to note that technology companies have taken steps to address these concerns and improve user privacy by providing more transparency, enhancing privacy settings, and allowing users to delete voice recordings. However, these incidents highlight the need for users to be vigilant about their privacy settings and the potential risks associated with voice-activated devices. Users should also be aware of the data collection and storage practices of the virtual assistants they use and make informed decisions about their usage.

AI profiling

A third aspect of privacy in the AI environment is the right to object to automatic profiling by limiting the ability of commercial or government entities to combine personal data with big data amassed from other people to construct behavioral profiles using AI and machine learning.[374] AI tools are used to look for patterns in human behaviour. Having access to the correct data sets can be used to make inferences about everyday things that are deeply private and personal, such as how many residents of a neighbourhood are likely to visit a specific place of worship, what television programmes they might enjoy, and even roughly their sleeping patterns.

The use of AI techniques can identify groups, such as those who share a specific political or personal stance, and draw broad conclusions about individuals, including about their mental and physical health. Despite their probabilistic character, judgments and predictions provided by AI can often serve as the foundation for decisions that have an impact on people's fundamental rights. These issues are exacerbated in the context of the Judiciary, for example when judges rely on making decisions using the help of AI systems.[375]

The story of how Target used data analytics to predict that a teenage girl was pregnant before her family knew is a well-known example of the power of data analysis and predictive modeling in retail. Here's a summary of the case:

In 2012, an article in The New York Times revealed that Target, a US retail giant, had developed an algorithm to predict customers' shopping habits and preferences. They used this data to send targeted advertisements and coupons to customers. One of the most famous examples from this article involved a teenage girl.

Target's algorithm had identified that a teenage girl was buying unscented lotion, dietary supplements, and cotton balls. While these purchases might seem unrelated, the algorithm recognized that this combination of products was often indicative of pregnancy. The algorithm assigned a "pregnancy prediction" score to each customer based on their purchase history. Once the system had a high prediction score for

374 Ibid.
375 UN Human Rights Council (2021). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, availableat:https://www.ohchr.org/en/calls-for-input/2021/right-privacy-digital-age-report-2021

a customer, it would start sending them advertisements and coupons related to pregnancy and baby products. In this specific case, Target began sending the teenage girl coupons for baby products like diapers, cribs, and baby clothes.

The girl's father was shocked to find these pregnancy-related advertisements addressed to his daughter. He called the store to complain about the inappropriate ads he thought were being sent to his teenage daughter. However, a few days later, he discovered that his daughter was indeed pregnant.

The algorithm had accurately predicted the girl's pregnancy based on her shopping patterns, even before her family was aware of it. The combination of seemingly unrelated products in her purchase history, such as unscented lotion and cotton balls, indicated a high likelihood of pregnancy.

This case illustrates how advanced data analytics and predictive modeling can be used by retailers to understand customer behavior and send highly targeted advertisements. While this can be effective for marketing purposes, it also raises important questions about privacy and the ethics of collecting and using customer data. It's essential for companies to handle customer data responsibly and transparently to maintain trust with their customers.

Source: Duhigg C. (2012). How Companies Learn Your Secrets, available at: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

AI tools can also be used for profiling of judges. An interesting regulatory initiative that aims to safeguard judge's integrity and prevent profiling by AI is the French law on Programming and Reform of Justice (2019-2022). In its article 33, this regulation aims to prevent anyone - but especially legal technology companies focused on predicting and analyzing litigation - from publicly disclosing the behavior pattern of judges in relation to judicial decisions. It reads as follows, "The identity data of judges and members of the Judiciary may not be reused for the purpose or effect of evaluating, analyzing, comparing or predicting their actual or alleged professional practices."[376]

For facts of the case Nubian Rights Forum and others v. The Attorney General, Kenya, 2021 and to discuss its implications on digital surveillance and privacy in Kenya please read Privacy International's article **"Data Protection Impact Assessments and ID systems: the 2021 Kenyan ruling on Huduma Namba"**

Many governments have started digitalizing their public services, bringing them online, and offering national digital identification (ID) systems. By amassing big volumes of personal data, these digital systems and databases threaten the right to privacy of citizens. National digital identity programs are just one of the many examples of how digital rights can be violated by governments. These programs require collecting and storing sensitive personal data and biometric identifiers to create a single digital

376 French Law on Programming and Reform of Justice (2019-222), available at: https://www.wipo.int/wipolex/en/legislation/details/18789

ID, to improve the delivery of government services. However, it is important for governments to understand the potential risks to users before creating centralized databases of personal and biometric data. To prevent human rights violations and cybersecurity, laws must include proper protections before rolling out such programs. Many national and regional courts have acted upon lawsuits against these digital systems brought by citizens and civil society organisations (CSOs).

One such case is Nubian Rights Forum and others v. The Attorney General, Kenya, 2021, where the High Court of Kenya declared unconstitutional the country's National Integrated Identity Management System (NIIMS), a digital ID system.[377] The Court stated that a Data Protection Impact Assessment should have preceded the program and that an appropriate legal framework to mitigate privacy and data protection risks should have been in place before the implementation of the NIIMS.[378] This passage highlights the common pitfalls that court rulings in various countries have identified when deciding challenges brough about by CSOs and other stakeholders against digital ID systems. In another case, the Mauritian Supreme Court emphasized the lack of adequate defense against security risks associated with biometrics. The Aadhaar judgement in India expressed concerns about centralized databases, while the Supreme Court of the Philippines identified the risk of individual tracking through a national identity system. Finally, the Kenyan High Court identified the risk of exclusion due to biometric and other identity system registration failures.[379]

Activity: Targeted advertising and price discrimination propelled by AI algorithms. Training participants discuss the main legal issues and human rights impacted by targeted advertising and personalised pricing. What laws are applicable under these circumstances?

Targeted advertising

In today's digital age, self-learning algorithms have become an integral part of big data analytics. With the help of AI, private companies can collect a plethora of personal information, such as your browsing habits, social media likes, health data, and purchasing patterns. These details can then be used to create a detailed profile of an individual, which can be further utilized for online tracking and profiling. This helps companies to tailor their advertising, pricing, and contract terms to the customer's specific profile, and leverage the consumer's biases and willingness to pay, all thanks to the findings of behavioral economics. Additionally, AI-based insights can also be used for scoring systems, which can decide whether a specific consumer is eligible to purchase a product or take up a particular service. Using self-learning algorithms in big data analytics allows private companies to gain a detailed insight into one's personal circumstances, behavior patterns, and personality (purchases, sites visited, likes on

377 See: https://globalfreedomofexpression.columbia.edu/cases/nubian-rights-forum-v-attorney-general.
378 UNESCO (2022). Guidelines for judicial actors on privacy and data protection, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381298
379 Privacy International (2022). Data Protection Impact Assessments and ID systems: the 2021 Kenyan ruling on Huduma Namba, available at: https://privacyinternational.org/news-analysis/4778/data-protection-impact-assessments-and-id-systems-2021-kenyan-ruling-huduma

social networks, health data). AI is used in online tracking and profiling of individuals whose browsing habits are collected by "cookies" and digital fingerprinting and then combined with queries through search engines or virtual assistants. Companies can tailor their advertising, prices, and contract terms to the respective customer profile and – drawing on the findings of behavioral economics – exploit the consumer's biases and/or her willingness to pay. AI-based insights can also be used for scoring systems to decide whether a specific consumer can purchase a product or take up a service.

The increasing use of targeted advertising, which relies on internet tracking and profiling, has raised concerns about privacy and data protection. With everything being automated, users are often left unable to give meaningful consent. The use of AI for intensive data processing may further exacerbate other rights violations, particularly in cases where personal data is used to target individuals in contexts such as insurance or employment applications. In some cases, algorithms can even pose a threat to both the right to privacy and freedom of expression. This creates growing issues for privacy and data protection. Targeted advertising uses internet tracking and profiling based on the person's expected interests. All these methods have incapacitated users from giving meaningful consent because everything is automated. Intensive data processing using AI may exacerbate other rights violations when personal data is used to target individuals, such as in the context of insurance or employment applications, or when algorithms threaten both the right to privacy and the freedom of expression.[380] For instance, social media algorithms decide the content of a user's newsfeed and influence the number of people who see and share information. Search engine algorithms index content and determine what appears at the top of search results. These algorithms threaten media pluralism and suppress the diversity of viewpoints.[381] To illustrate this, in 2023, Meta was fined €390 million by the Irish Data Protection Committee for violating the GDPR. The regulator has alleged that Meta's use of personal data on Facebook and Instagram, specifically for personalized advertising, did not comply with the GDPR.[382]

Price discrimination

In the digital age, AI plays a significant role in helping businesses tailor their offerings to individual customers. By analyzing consumer behavior and preferences, AI algorithms can estimate the highest price point that a particular customer is willing or able to pay. This approach is particularly relevant for industries such as credit and insurance, which operate on risk-based cost structures that take into account the unique features of each consumer. However, the question of whether regulators should allow price discrimination in other sectors based on a customer's ability to pay is a complex and contentious issue that requires further exploration and debate.AI supports digital businesses in presenting consumers with individualised prices, and offering to each consumer an approximation of the highest price point that consumer may be able or willing to pay. Certain markets, such as credit or insurance, operate on cost structures

380 Council of Europe (2017). Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, available at: https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a

381 Access Now (2018). Human rights in the age of artificial intelligence, available at: https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdf

382 The Data Protection Commission (2023). Data Protection Commission announces conclusion of two inquiries into Meta Ireland, available at: https://www.dataprotection.ie/en/news-media/data-protection-commission-announces-conclusion-two-inquiries-meta-ireland

based on risk profiles correlated with features distinctive to individual consumers, suggesting that it may be reasonable to offer different prices (e.g., interest rates) to different consumers. Should regulators allow price discrimination in other cases, too, based on the ability of different consumers to pay?[383]

It is concerning that consumers are typically unaware when advertising, information, prices, or contract terms have been personalized based on their profile. If an algorithm calculates a certain score that results in a contract not being offered or only being offered at unfavorable conditions, consumers often struggle to comprehend how this score was generated. Moreover, the complexity, unpredictability, and semi-autonomous behavior of AI systems can pose challenges for enforcing consumer legislation, as it is difficult to trace decisions back to a single actor and ensure legal compliance. Consumers are not usually aware that advertising, information, prices or contract terms have been personalized according to their profile. Suppose a certain contract is not concluded or only offered at unfavorable conditions because of a certain score calculated by an algorithm. In that case, consumers are often unable to understand how this score was achieved. Complexity, unpredictability, and semi-autonomous behavior of AI systems can also make effective enforcement of consumer legislation difficult, as the decision cannot be traced to a singular actor and therefore cannot be checked for legal compliance.

All these practices of automated profiling enabled by AI have had severe implications for the enjoyment of the right to private and family life. The trails of personal information, such as the digital exhaust knowingly or unknowingly produced by cell phones, computers, and other technologies, left in the digital realm are never-ending. How that personal information is collected and used by third parties is a huge concern for regulators.[384]

AI is used in online tracking and profiling of individuals whose browsing habits are collected by "cookies" and digital fingerprinting, and then combined with queries through search engines or virtual assistants. Mobile apps process behavioural data (such as location and health data) from smart devices. This creates growing issues for privacy and data protection. Targeted advertising uses internet tracking and profiling based on the person's expected interests. The use of all these

For a firsthand experience of online tracking, training participants should go online to Google's Ads Preference manager at: http://www.google.com/ads/preferences/ and look at markers used by the company to define them and assess how accurate these are.
The information tracked is used to create digital profiles of users to which access is sold in the market place, including specialized exchanges, to help advertisers market their products better

383 European Parliament (2019). Artificial Intelligence: Challenges for EU Citizens and Consumers, available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/631043/IPOL_BRI(2019)631043_EN.pdf
384 Perry W. L., McInnis B., Price C. C., Smith S., Hollywood J. S. (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, RAND Corporation: Santa Monica, available at: https://www.rand.org/pubs/research_reports/RR233.html

methods has incapacitated users to give meaningful consent because everything is automated. Even though users may be asked for consent as required by law (a) they do not always necessarily understand what they are being asked of; (b) still the terminology and terms and conditions may be confusing and running into many pages; and (c) with so much content online, users suffer from information overload.

Case Study: Case law on profiling people through ADM

In 2018, the Italian Data Protection Authority (Garante) discovered that a data controller was violating the national data protection law by offering personalized rates to customers of its car-sharing service based on their observed habits and characteristics. In the administrative procedure, the defendant disputed, claiming that there was no "categorization" of the service's users because the information used to determine the fees was not linked to the subjects. The DPA rejected the defendant's objections, finding that it was evident that there had been personal data processing in this case, that it was exclusively automated processing, and that it was intended to define a person's profile or personality or to analyze their habits or consumption choices. The Italian Supreme Court (Corte Suprema di Cassazione) upheld this decision in November 2021, which resulted in an administrative fine of 60.000 EUR. In the appeals process, the Supreme Court sided with the Garante because it ruled that processing personal data using an algorithm to determine an individual rate constitutes profiling, even if the data is neither stored by the controller nor attributable to the data subject.

Source: Future of Privacy Forum (2022). GDPR and the AI Act interplay: Lessons from FPF's ADM Case Law Report, available at: https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report

Data Anonymization does not always lead to privacy protection

The privacy of data is usually protected through anonymization. Identifiable aspects such as names, phone numbers, and email addresses are stripped out. Datasets are altered to be less precise, and "noise" is introduced to the data. However, a study published by Nature Communications suggests that anonymization does not always protect privacy. Researchers have developed an ML model that estimates how individuals can be re-identified from an anonymized data set by entering their zip code, gender, and date of birth.

Source: Rocher L., Hendrickx J. M., de Montjoye Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models, Nature Communications, 10 (3069), available at: https://www.nature.com/articles/s41467-019-10933-3

Emerging privacy issues

The creation of new data is a unique challenge in the automated processing of personal data. It is often possible for personal data to be combined, leading to the creation of second and even third generations of data about a particular person. When compared to a much bigger data set, two seemingly unrelated pieces of information might "breed" and produce new data, unbeknownst to the data subject. Significant questions are raised regarding the concepts of consent, openness, and personal autonomy.[385] Issues that deserve further attention: How much control will subjects have over the information collected on them? Given their stake in the provision of personal data for ML training purposes, should individuals have the right to utilize the model or at least know what it is used for? Could data-seeking ML systems inadvertently violate people's privacy if, for instance, analyzing the genome of one family member revealed health data about other family members?[386]

Discussion point (10-15 minutes): "The power of privacy (1/5): Does the internet know where you live?"

Training participants watch the video produced by the Guardian, "The power of privacy (1/5): Does the internet know where you live?" and discuss how the notion of privacy has changed in the digital realm and how this has impacted their work. They also discuss examples from their respective jurisdictions.

Source: https://www.youtube.com/watch?v=iA89GhyLao8

385 Committee of Experts on Internet Intermediaries (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, available at: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

386 European Parliament (2020). The ethics of artificial intelligence: Issues and initiatives, available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)634452

All these challenges have been exacerbated in public sector settings. According to the NGO Access Now, with the expansion of the Internet and the growth of new technologies, government surveillance has increased, and AI is enabling more intrusive surveillance capabilities than ever before. Even though no completely centralized government facial recognition system is currently known to exist, some countries have tried to deploy more CCTV cameras in public areas and centralize their facial recognition systems.[387] Half of all US adults are now in law enforcement facial recognition databases.[388] The use of these technologies poses a threat to anonymity, and the dread of being observed can prevent the exercise of other rights, such as the freedom of association. The underprivileged demographics, who are already under the frequent control of the security forces, would experience the negative effects of AI-powered surveillance the most directly. In addition, since monitoring the entire population 24 hours a day, seven days a week is neither essential nor proportional to the purpose of public safety or crime prevention, it would almost likely violate the right to privacy.[389]

387 AccessNow (2018) AI and human rights, available at: https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdf
388 Ibid.
389 Ibid.

Case studies

The case of the SARI Real Time facial recognition system, Italy

The Italian data protection authority (DPA) has published an opinion on the Sari Real Time system presented for review by the country's Ministry of Interior, claiming that if employed as intended, the technology "would establish a type of mass monitoring." Sari, which is not yet operational, is a facial recognition system that, using several cameras set in specified geographical areas, would analyse the faces of individuals filmed in real-time and compare them to a prepared database of up to 10,000 faces. Sari would be implemented "where there is a need for a facial recognition technology to help police forces in the management of order and public safety, or in response to the unique requirements of the judicial police."

The DPA has stated that Sari 'would conduct a large-scale automated processing that might include those present at political and social demonstrations who are not the subject of police "attention"' In addition, the fact that "the identification of a person would be accomplished through the processing of the biometric data of all persons present in the monitored space" would result in a "transition from targeted surveillance of specific individuals to the prospect of universal surveillance". The DPA determined that the Ministry had not clarified the legal foundation upon which it would conduct such actions. It was stated that "an effective regulatory framework should take into account all the rights and freedoms involved and identify the scenarios in which the use of such systems is permissible, without giving a great deal of discretion to the users."

Source: DigWatch, Italian data protection authority: Sari facial recognition system proposed by Ministry of Interior could lead to mass surveillance, available at: https://dig.watch/updates/italian-data-protection-authority-sari-facial-recognition-system-proposed-ministry-interior

Use of live facial recognition technology in Buenos Aires, Argentina

Between 2019 and 2022, live facial recognition technology was implemented in Buenos Aires, Argentina's capital city, to aid security forces in identifying potential criminals who were wanted in the country's national fugitive database. The system relied on live footage from video monitoring systems stationed throughout the city, including the three main railway stations, and the underground transport network, which is used by over 1.3 million passengers each day. However, in April 2022, a court order was passed to temporarily suspend the use of the technology due to allegations of unauthorized searches. And in September 2022, a city court ruled that the current conditions under which the system was operating were unconstitutional, which is expected to extend the suspension of the facial recognition system further. According to Argentina's Association for Civil Rights (Asociación por los Derechos Civiles — ADC), facial recognition technology has been implemented not only in the capital but also in other regions including the provinces of Córdoba, Salta, and Mendoza, as well as in the county of Tigre in Buenos Aires. It has been reported that there are also plans to deploy the technology in the province of Santa Fe. This information was accurate as of early 2021.

Use of facial recognition technology in Brazil

The use of facial recognition technology is quite widespread in Brazil, with deployments identified in 30 cities as of 2019. This technology is employed for a variety of purposes, including preventing fraud in the distribution of social benefits. It has been used to verify the identities of beneficiaries of public transport subsidies in numerous Brazilian cities and track school attendance requirements for cash transfer programs in the state of Pernambuco. However, facial recognition technology has also been deployed for marketing purposes, such as placing advertisements in front of passengers in the São Paulo Metro using highly controversial emotion detection techniques. This project was eventually rolled back after a local court declared that data collection on Metro passengers did not meet minimum consent requirements.

Argentina and Brazil are both federal systems that have a complex coexistence of municipal, state, and federal laws. This often leads to a patchwork of regulations with varying standards and safeguards that can be quite confusing. This complexity has led to challenges in justifying the legality of facial recognition deployments. In Argentina and Brazil, local governments have implemented a mix of city legislation and state-level regulatory proposals that often fall short of the standards outlined in their respective constitutions, international human rights treaties, and federal laws.

Source: Chatam House (2022). Regulating facial recognition in Latin America, Policy lessons from police surveillance in Buenos Aires and São Paulo, available at: https://www.chathamhouse.org/2022/11/regulating-facial-recognition-latin-america/03-facial-recognition-rollouts-trends-buenos

Classification of data protection as an independent right

The classification of data protection as an independent right has been a point of contention in international courts and academia. It stems from the fact that data protection, as a regulatory issue, arose in part from privacy regulations, norms, and concerns, and evolved into new sets of obligations imposed on public authorities and commercial entities to provide individuals with control over the information that concerns them, as well as the means to achieve that control — access to this information, confirmation of its existence, correction of incorrect data, etc.

However, data protection extends beyond privacy concerns. There may be important data protection concerns when privacy considerations are irrelevant or secondary, as illustrated below in the section that deals with

data protection principles.[390] Data protection builds on the right to privacy, but also encompasses other data subject rights vis-à-vis the government and large corporations that collect, process, and store personal data, such as the right to be informed, the right of access to personal data, the right to be forgotten, the right to rectification, the right to data portability, the right to object to processing, and rights related to automated decision-making and profiling.[391]

Many countries around the globe recognize data protection as a fundamental right. Personal data protection is incorporated as an independent right in various statutes, including the European Union's Charter of Fundamental Rights (Article 8). It was also recently acknowledged as such by the Brazilian Supreme Court. Similarly, in a recent case (Justice K. S. Puttaswamy (Retd.) v. Union of India [392]), the Indian Supreme Court affirmed privacy as a fundamental right.[393]

Data protection rights related to automated decision-making and profiling

In many jurisdictions, data subjects have rights related to automated decision-making and profiling. This covers various profiling techniques, which may involve assessing specific personal characteristics linked to an individual that evaluate or forecast behaviour related to performance at work, financial condition, health, personal preferences, hobbies, reliability, conduct, or location. The right to be exempt from automated decision-making is usually guaranteed to data subjects where those decisions have a material impact on their lives. However, these rights do not apply to partially automated decisions. Nor do they necessarily ensure that, in practice, an affected individual can readily detect whether they have been treated unequally vis-à-vis others, and if so, whether such differential treatment amounted to discrimination and was thus unlawful. The data subject has freedom to waive some of their rights by consenting to specific practices that would otherwise constitute a rights violation, thereby forgoing the protections these rights provide.

For example, there is a significant risk that data protection rights would be too readily waived by individual right-holders in a networked age built upon a 'free services' business model: in return for 'free' access to digital services and the efficiency and convenience they offer, individuals will willingly exchange their personal data.[394] On the other hand, the core data protection principles include mandatory obligations imposed on data controllers that cannot be waived by individual right-holders, including the principles of lawfulness of

390 Ibid.
391 Ibid.
392 Status as Fundamental Right (2017). Justice K.S. Puttaswamy (Retd.) v. Union of India, available at: https://privacylibrary.ccgnlud.org/case/justice-ks-puttaswamy-ors-vs-union-of-india-ors
393 UNESCO (2018). Legal Standards on Freedom of Expression, Toolkit for the Judiciary in Africa, available at: https://unesdoc.unesco.org/ark:/48223/pf0000366340
394 Committee of Experts on Internet Intermediaries (MSI-AUT) (2019). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework, Council of Europe Study, DGI/2019/05, available at: https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168096bdab

the processing, of purpose specification and data minimization. This offers more systematic and robust protection of the core underlying values and collective interests that data protection regimes ultimately seek to protect.[395]

395 Ibid.

Activity: Training participants read the case studies below and discuss how data protection laws are enforced in their jurisdictions, noting renowned cases and comparing them to the GDPR cases below. How would a similar case be judged and decided in your jurisdiction? What laws would apply?

Meta's privacy violations in the EU

After it was discovered that Facebook users' personal information was posted on an online hacker forum, Meta, the owner of Facebook, was fined €265 million by the Irish regulator, the Data Protection Commission, for violating data protection laws. The leaked information included the complete names, contact information, dates of birth, and localities of Facebook users in 2018 and 2019.

Meta acknowledged that the information had been scraped using technologies that were intended to assist individuals in discovering friends by phone numbers. Facebook was penalized for "failure to apply Data Protection by Design and Default" in accordance with the GDPR. The fine may have been avoided if this feature had been designed to be more secure.

Source: Satariano A. (2022). Meta Fined $275 Million for Breaking E.U. Data Privacy Law, available at: https://www.nytimes.com/2022/11/28/business/meta-fine-eu-privacy.html

Google's privacy violations in the EU

On January 6, 2022, the French data protection authority (CNIL) fined Google Ireland €90 million. The fine pertains to how YouTube's cookie consent processes are implemented by Google Europe. The Google Ireland fine was one of two penalties issued in the same case; the other was made against Google LLC of California (which operates Google Search).

Google should have enabled YouTube users to reject cookies easily, according to the CNIL. YouTube places cookies on devices for marketing purposes to track online activities. It is simple to accept cookies on YouTube but more difficult to reject them. The CNIL observed that rejecting cookies needed many clicks, but accepting cookies required just one. Under GDPR, consent must be "voluntary": If an offer can be accepted with a single click, it should also be possible to reject it with a single click.

The CNIL justified the comparatively hefty punishment by citing the great number of YouTube users and Google's enormous earnings from the site.

Source: Lomas N. (2022). France spanks Google $170M, Facebook $68M over cookie consent dark patterns, available at: https://techcrunch.com/2022/01/06/cnil-facebook-google-cookie-consent-eprivacy-breaches/

Legitimate restrictions to the right to privacy

The ICCPR (Article 2) mandates that states, parties to the ICCPR, "respect and ensure" without discrimination the rights enumerated in the Covenant for all individuals within their territory and under their jurisdiction. However, privacy rights are not absolute. In many jurisdictions, law enforcement agencies are exempt from data privacy legislation[396]. Governments may legitimately disrupt a person's privacy under certain circumstances specified by the law, such as emergencies or threats to national security. Any limitations on the rights enumerated in the ICCPR must be permissible under the relevant ICCPR provisions. Governments must justify their surveillance actions and demonstrate that any invasion of privacy is established in laws and regulations that are clear and precise, necessary[397] for achieving legitimate government objectives, and proportional to achieving these limited objectives. An independent, unbiased, and competent judicial or administrative institution must oversee surveillance actions by law enforcement agencies. Moreover, government officials and others must be held accountable for misconduct and errors.[398]

According to the Office of the United Nations High Commissioner for Human Rights, state surveillance activities must follow the law. The exceptions to digital surveillance should be limited and based on the principles of necessity and proportionality to ensure adequate data privacy across all government branches.[399] The following minimum requirements should govern the enactment of surveillance-specific laws:

- The law must be accessible to the public and adequately specific. The law needs to precisely define the scope of surveillance discretion granted to the government agency and the manner of surveillance. The law should also describe the nature of the offense and the class of individuals who may be subject to surveillance. Unspecific references to "national security" or "public health" do not qualify as specific and legitimate justifications as they are vague and broad. Surveillance must be founded on reasonable suspicion, and any authorizing decision for surveillance must be sufficiently targeted. The law must precisely define the competencies of the institution with authority to conduct digital surveillance.
- Regarding its scope, the legal framework for surveillance should also include surveillance requests from the government to businesses. The legal framework should also include access to information held extraterritorially and the exchange of information with other states. The

396 UNESCO (2022). Guidelines for Judicial Actors on Privacy and Data Protection, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381298?posInSet=1&queryId=26b4d8e4-fbc6-4c96-82ea-3795782bae8a

397 The necessity component of the test for restrictions is the most challenging and litigated. It involves various factors in various international jurisdictions. Two key factors in determining necessity are (i) the restriction must serve an urgent social need, and (ii) the justifications for the restriction must be sufficient and pertinent. See: Icelandic Human Rights Centre, https://www.humanrights.is/en/human-rights-education-project/comparative-analysis-of-selected-case-law-achpr-iachr-echr-hrc/the-right-to-freedom-of-opinion-and-expression/permissible-limitations. See also: Australian Human Rights Commission, Permissible Limitations on Rights, https://humanrights.gov.au/our-work/rights-and-freedoms/permissible-limitations-rights

398 Icelandic Human Rights Centre, https://www.humanrights.is/en/human-rights-education-project/comparative-analysis-of-selected-case-law-achpr-iachr-echr-hrc/the-right-to-freedom-of-opinion-and-expression/permissible-limitations. See also: UN (2018). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, available at: https://documents.un.org/doc/undoc/gen/g18/239/58/pdf/g1823958.pdf?OpenElement

399 UN Human Rights Council (2018). The right to privacy in the digital age. Report of the United Nations High Commissioner for Human Rights, availableat:https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/239/58/PDF/G1823958.pdf?OpenElement

law must explicitly establish a structure to ensure accountability and transparency within government organizations conducting surveillance.

- Surveillance powers can only be justified if they are strictly necessary for attaining a legitimate goal and if they satisfy the requirement of proportionality. The scope of surveillance must be limited to preventing or investigating the most serious offenses or threats. The duration of the surveillance should be kept to the absolute minimum required to achieve the specified objective. Based on strict necessity and proportionality, the law should contain strict rules for using and storing the collected data, and define precisely the circumstances under which the collected and stored data must be erased. The same rules of legality, strict necessity, and proportionality must apply to the exchange of intelligence.[400]
- When governments contemplate targeted hacking, they should proceed with extreme caution, resorting to such measures only in exceptional circumstances, for the investigation or prevention of the gravest offenses or threats, and with the participation of the Judiciary. The design of hacking operations should be limited, limiting access to specific targets and categories of information. States should not compel private entities to assist in hacking operations, as doing so would compromise the security of their own products and services. Compulsory decryption may be allowed only on a case-by-case basis, with a warrant and the preservation of due process rights.[401]

Surveillance measures, such as requests for communications data from businesses and intelligence sharing, should be authorized, reviewed, and supervised by independent bodies at all stages, including when they are initially ordered, while they are being carried out, and when they are terminated.[402] The independent body authorizing particular surveillance measures, preferably a judicial authority, must ensure that there is sufficient evidence of a threat and that the proposed surveillance is targeted, strictly necessary, and proportionate before authorizing (or rejecting) the surveillance measures ex ante.

Oversight frameworks include administrative, judicial, and/or parliamentary agencies. The oversight bodies should be independent of the surveillance authorities and endowed with the necessary expertise, skills, and resources. Institutionally, the rules should differentiate between and separate the functions of authorization and oversight. In addition to periodic evaluations of surveillance capabilities and technological advancements, independent supervision bodies should investigate and monitor the activities of those conducting surveillance

400 Human Rights Council (2013). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, available at: https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A.HRC.23.40_EN.pdf

401 UN Human Rights Council (2015). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression,availableat:https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/095/85/PDF/G1509585.pdf?OpenElement

402 International Covenant on Civil and Political Rights (2015). Concluding observations on the fifth periodic report of France, available at: https://tbinternet.ohchr.org/_layouts/15/TreatyBodyExternal/Download.aspx?symbolno=CCPR%2FC%2FFRA%2FCO%2F5&Lang=en

and accessing its products.[403] Agencies conducting surveillance should be required to provide all the information necessary for effective oversight upon request, submit regular reports to the oversight bodies, and maintain records of all surveillance measures. Additionally, oversight processes must be open and subject to appropriate public scrutiny, and oversight bodies' decisions must be subject to appeal or independent review.[404]

Principle of transparency: Open discussion and scrutiny are crucial to comprehending the benefits and constraints of surveillance techniques, therefore, state authorities and oversight bodies should also engage in public information about the existing laws, policies, and practices in surveillance and communications interception, as well as other forms of processing personal data.[405] The surveillance agency should explain the limitation on the right to privacy to those who were the target of surveillance. Moreover, those subjected to surveillance should have the right to change and remove unnecessary personal information if it is no longer required for ongoing or future investigations.[406]

In principle, to be legal, restrictions on the right to privacy through the national human rights framework, data protection, cybersecurity, cybercrime, and digital surveillance or ICT laws and policies must meet certain minimum international human rights law standards. These standards can be found in the UN General Assembly Resolution on the Right to Privacy in the Digital Age of 2014[407], the 2014 Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism[408], and the Report of the Office of the United Nations High Commissioner for Human Right to Privacy in the Digital Age.[409] According to these standards[410], to be legal, the restrictions on the right to privacy made by governments should be:

- **Imposed only for protecting legitimate purposes:** With respect to the right to privacy, digital surveillance should only be authorized in pursuit of the most vital national goals. In addition, it must be demonstrated that the restriction put on the right (such as an invasion of privacy to safeguard national security or the right to life of others) can reasonably accomplish its intended objective. The burden is on the authorities attempting to restrict the right to demonstrate that the restriction serves a legitimate aim.
- **Lawful:** Limits on the right to privacy must be stated clearly and

403 See European Court of Human Rights, Kennedy v. United Kingdom, application No. 26839/05, judgment of 18 May 2010.
404 https://www.cipil.law.cam.ac.uk/projects/human-rights-big-data-and-technology-hrbdt-project
405 UN Human Rights Council (2009). Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, available at: https://digitallibrary.un.org/record/648291?ln=fr
406 UN Human Rights Council (2017). Report of the Special Rapporteur on the right to privacy, available at: https://digitallibrary.un.org/record/3845912?ln=en&v=pdf
407 UN Human Rights Council (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, available at: https://digitallibrary.un.org/record/777869?v=pdf
408 See: https://www.ohchr.org/en/special-procedures/sr-terrorism
409 UN Human Rights Council (2021). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, available at: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx
410 It has to be noted that these standards are not universally accepted by all governments. Many governments interpret the provisions of the ICCPR differently. For instance, the United States has historically noted (see page 235, available at: https://2017-2021.state.gov/wp-content/uploads/2019/10/2018-Digest-Final-Draft.pdf#page=235) that Article 19 of the ICCPR does not impose a standard of legality, necessity, and proportionality—only that surveillance cannot be unlawful of arbitrary.

unambiguously in the law and should be reviewed frequently to ensure that privacy protections and safeguards keep pace with the rapid digital technology developments.[411] According to the Report of the Office of the United Nations High Commissioner for Human Rights, "The right to privacy in the digital age": "interference that is permissible under national law may nonetheless be "unlawful" if that national law is in conflict with the provisions of the International Covenant on Civil and Political Rights".[412]

- ➔ **Compliant with the principle of non-discrimination in their design and application:** Limits on the right to privacy should not discriminate against any vulnerable groups.
- ➔ **Necessary and proportionate:** Digital surveillance is a very intrusive act that violates the right to privacy. Prior approval from a competent judicial authority is necessary for proportionate digital surveillance. The restriction must be essential for achieving a legitimate aim, proportional to the objective, and the least invasive surveillance methods shall be used.[413]

Governments limit the right to privacy because of the following reasons:

- National security
- Public safety
- National economic well-being
- Protection of the rights and freedoms of others.
- Prevention of disorder or crime
- Protection of health or morals[414]

411 MISA Zimbabwe, Konrad Adenauer Stiftung (2021). Cybersecurity and Cybercrime Laws in the SADC Region: Implications on Human Rights, available at: https://fdocuments.net/document/cybersecurity-and-cybercrime-laws-in-the-sadc-region.html?page=3

412 UN Human Rights Council (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, available at: https://digitallibrary.un.org/record/777869?v=pdf

413 International Commission of Jurists, Regulation of Communications Surveillance and Access to Internet in Selected African States, available at: https://www.kas.de/documents/275350/0/Report-on-Regulation-of-Communications-Surveillance-and-Access-to-Internet-in-Selected-African-States.pdf/66dbd47d-4d7d-2779-a595-a34e9f93cfbb?t=1639140695434

414 See: https://africaninternetrights.org/en/node/2558#:~:text=This%20advocacy%20toolkit%20provides%20an%20overview%20of%20the,the%20formulation%20and%20implementation%20of%20data%20protection%20frameworks.

In depth: Human rights-based approach (HRBA) to assessing the impact of regulation on the right to privacy in the digital environment

Countries that enact cybercrime, cybersecurity, and data protection laws should follow the HRBA to the drafting of digital regulation. An HRBA is based on the principles derived from international and regional treaties and places human rights at the center of all policy formulation and legislative drafting. This approach's fundamental elements are participation, accountability and transparency, non-discrimination and equality, rights holder empowerment, and legality. Digital surveillance regulation must be unambiguous regarding which agencies can conduct surveillance, who can judge requests to undertake surveillance, what legal tests a court must apply to requests, and what legal penalties apply to unauthorized surveillance.[415] Lawyers and advocacy groups and CSOs that work in the area of digital privacy should use the HBRA as a tool in assessing if the restrictions imposed on the right to privacy by the government are legitimate, lawful, compliant with the principle of non-discrimination in their design and application, and necessary and proportionate.[416]

An HBRA should entail an assessment of the national digital regulation against the International Principles on the Application of Human Rights to Communications Surveillance[417]. The figure below outlines the key areas that these Principles focus on:

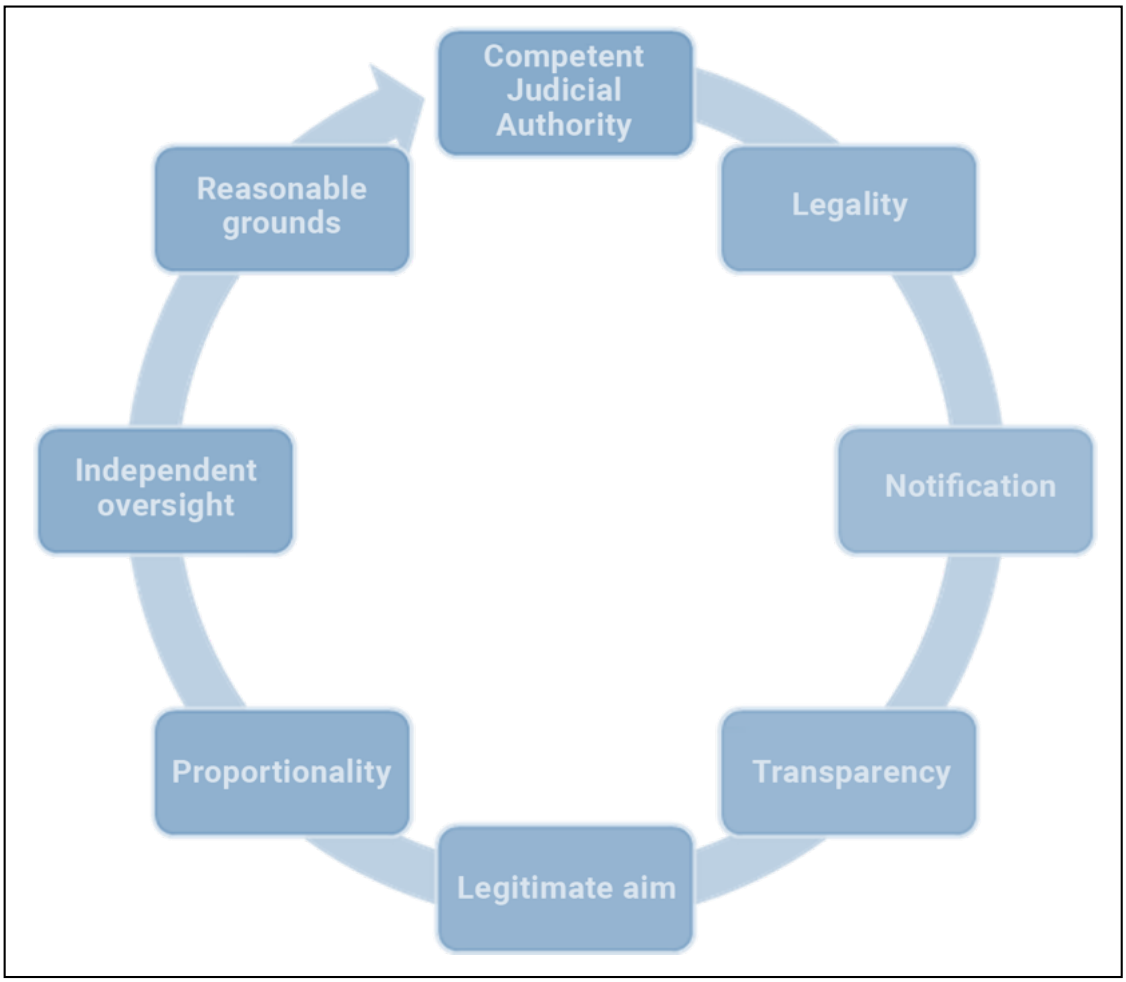
- ➔ Prior authorization of surveillance by a competent judicial authority: Is there a judge with expertise in digital technology and human rights who can evaluate and authorize surveillance requests from investigating government agencies?
- ➔ Legitimate aim: does the law establish certain lawful purposes of surveillance, such as prevention of terrorism or grave crime with a legal punishment of 10 or more years in jail?
- ➔ Reasonable grounds: Are judges empowered to determine if there is a high level of threat to a legitimate objective and a high likelihood that surveillance would generate evidence that eliminates the threat?
- ➔ Legality: Is surveillance carried out exclusively within the limitations and by the agencies specified by law? Does the law make any other surveillance illegal and stipulate penalties?
- ➔ Necessity: Are judges authorized to determine whether monitoring is required to secure the evidence and that no less intrusive method exists to achieve the legitimate purpose?
- ➔ Proportionality: Are judges empowered to determine whether the proposed surveillance is limited in scope and the duration is proportional to the evidence required to eliminate the threat?

415 See: https://unsdg.un.org/2030-agenda/universal-values/human-rights-based-approach

416 Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Institute of Development Studies: Brighton, available at: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/16893/Roberts_Surveillance_Law_in_Africa.pdf

417 Over 600 groups, including Privacy International, the Open Rights Group, the Electronic Frontier Foundation, and the Association for Progressive Communications, coordinated the writing of the International Principles, available at: https://www.eff.org/files/necessaryandproportionatefinal.pdf

- Subject notification: Does the law require that the subject of surveillance be advised of the surveillance as soon as possible to provide an opportunity for legal appeal and due process?
- Transparency reports: Do annual reports on openness make public the amount of surveillance requests, justifications, and authorizations?
- Independent oversight: Do surveillance practices have any public monitoring mechanisms to ensure their accountability and transparency?[418]



Source: Adapted from Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Brighton: Institute of Development Studies, available at: DOI: 10.19088/IDS.2021.059

418 Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Institute of Development Studies: Brighton, available at: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/16893/Roberts_Surveillance_Law_in_Africa.pdf

3. Approaches to AI governance

As AI rapidly integrates across all sectors, it is important for judicial operators to consider the unique benefits and risks associated with different AI systems. Virtual assistants, self-driving vehicles, and video recommendations for children all present varying levels of benefits and risks. Therefore, policy making and governance must be approached differently for each specific AI system depending on the risks involved, their severity, as well as their impact on human rights. Table 7 below gives an overview of the guiding principles in governing AI.

Table 7. Select guiding principles in governing AI

Principles	Key issues in implementing the principles
The greater the risk to human rights, the tougher the legal standards should be for the use of AI technology.	Sectors where the stakes for encroachment on individual fundamental rights are high, such as national security, criminal justice, law enforcement, health, and social protection should have priority. A risk-proportionate approach to AI regulation will necessitate the prohibition of specific AI technologies, applications, and use cases that produce potential or actual impacts that violate international human rights, including those that fail the necessity and proportionality requirements.[419]
AI applications that discriminate should not be permitted.	The social scoring of individuals by governments[420] or the use of AI systems that classify individuals into clusters based on prohibited discriminatory factors should be outlawed. Governments will need to control the use and procurement of AI technologies whose deployment in the Judiciary poses dangers to human rights. When human rights violations are likely to occur, the requirement of human monitoring (human in the loop) should be mandated. Governments should postpone the deployment of potentially high-risk technologies, such as remote real-time facial recognition, until it can be guaranteed that their implementation will not violate human rights.[421]
If an AI system is used to engage with humans in the context of public services, particularly justice, welfare, and healthcare, the user must be advised and informed of the option to consult a professional upon request and without delay.	Those who have had a decision made about them by a public authority that is solely or substantially based on the output of an AI system should be alerted and given the aforementioned information as soon as possible.[422] This may take the form of either public disclosure of information on the system in issue, its processes, direct and indirect effects on human rights, and actions taken to identify and mitigate adverse human rights consequences of the system, or an impartial, thorough, and effective audit. In every instance, the information provided should permit a meaningful evaluation of the AI system. No AI system should be so complicated that human evaluation and inspection are impossible. ADM systems that cannot be held to adequate transparency and accountability standards should not be utilized in public service delivery.[423]

419 The proposed AI Act of the European Union takes such a risk-based approach.

420 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, para. 75, available at: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da; see also: UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

421 European Parliament (2019). A governance framework for algorithmic accountability and transparency, available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262; also see: CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, available at: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da

422 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, available on: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da

423 Ibid.

The EU AI Act as an example of risk-based regulation of AI

In 2019, following the publication of the Ethics Guidelines for Trustworthy AI[424] the European Commission started a multi-pronged approach for regulating AI and addressing AI-related risks. In addition to the AI Act, the new and amended civil liability rules[425] act in conjunction with other current and planned data-related policies, such as the GDPR[426], the Digital Services Act[427], the Data Act[428], and the Cyber Resilience Act[429].

The EU AI Act sets horizontal standards for developing, commercializing, and using AI-powered products, services, and systems within the EU. It provides fundamental AI risk-based guidelines applicable across all industries and includes a "product safety framework" with four risk categories, specifying market entry rules and certification for High-Risk AI Systems through a mandatory CE-marking process. This compliance regime also covers datasets used for machine learning training, testing, and validation to ensure fair outcomes.

The EU AI Act employs a risk-based strategy with multiple enforcement mechanisms. Low-risk AI applications would be subject to a more lenient regulatory framework, while those with unacceptable risks would be banned. As risk increases, more stringent regulations apply. These vary from lighter external certification requirements throughout the application's life cycle to non-binding self-regulatory soft law impact evaluations combined with codes of conduct.

The regulatory framework defines four levels of risk in AI:

(i) Unacceptable risk. AI systems harmful to people's rights, safety, and livelihoods shall be outlawed, including social scoring systems used by governments and voice-activated toys that promote risky behavior.[430]

(ii) High risk. The initial proposal (2021) included (i) critical infrastructure (e.g., transport), that could put the life and health of citizens at risk; educational or vocational training that may determine the access to education and professional course of someone's life (e.g., scoring of exams; (iii) safety components of products (e.g., AI applications in robot assisted surgery; (iv) employment, management of workers and access to self-employment (e.g., resume sorting services for recruitment purposes); (v) essential private and public services (e.g., credit scoring denying citizens opportunity to obtain a loan);

424 European Commission (2019). Ethics guidelines for trustworthy AI, available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
425 European Commission (2022). New liability rules on products and AI to protect consumers and foster innovation, available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5807
426 European Commission (2021). Data Protection, available at: https://commission.europa.eu/law/law-topic/data-protection_en
427 European Commission. The Digital Services Act, available at: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en
428 European Commission (2023). Data Act: Commission welcomes political agreement on rules for a fair and innovative data economy, available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3491
429 European Commission (2022). Cyber Resilience Act , available at: https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act
430 See: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

(vi) law enforcement activities interfering with human rights (e.g., evaluation of admissibility of evidence; (vii) migration, asylum and border control management (e.g. verification of authenticity of travel documents); (viii) administration of justice and democratic processes (e.g., applying the law to a concrete set of facts).

The proposal in December 2022, removed deepfake detection by law enforcement, crime analytics, and verification of the authenticity of travel documents from the list of high-risk AI systems. The Act does not encompass AI for national security, defense, and military purposes.

All remote biometric identification technologies are subject to tight regulations and are regarded as high-risk. In general, it is forbidden to employ remote biometric identification for law enforcement in areas open to the public. Only a few situations can be allowed as exceptions, such as when it is imperative to find a missing child, stop a specific and impending terrorist threat, or find, identify, or prosecute a perpetrator or suspect of a major crime. Such use is subject to proper time, location, and database search limitations, as well as approval by a judicial or other impartial body.[431]

(iii) Limited risk. AI systems with limited risk must adhere to specific disclosure requirements. Users should be aware that they are engaging with a machine when using AI systems like chatbots so they may decide for themselves whether to move forward or back.[432]

(iv) Minimal or no risk. Applications like spam filters or video games with AI are included in this.

Users assure human control and monitoring once an AI system is put on the market, while providers have a post-market monitoring structure in place. Authorities are in charge of market monitoring. Serious events and malfunctions will be reported by both providers and users.[433]

431 Ibid.
432 Ibid.
433 Ibid.

Human rights-based approaches to AI governance

A human rights-based approach is essential to build trustworthy AI systems in public service delivery. To ensure a rights-based approach in public sector operations, developing countries' governments should have a readily accessible analytical framework to assist them in identifying when AI components might impact human rights and how algorithmic accountability could mitigate those risks. Where AI systems threaten fundamental rights, countries should protect and promote those rights and ensure that private sector actors conduct due diligence and human rights impact assessments (HRIAs) according to their responsibility. The outcome of HRIAs should lead to different safeguards assigned to the specific risks and impacts established in the process.[434]

Governments around the globe, such as the United States (Blueprint for an AI Bill of Rights),[435] have attempted to address AI accountability and transparency issues through a human rights perspective. A valuable framework for conducting algorithmic impact assessments based on the human rights approach is provided by the fundamental rights and algorithm impact assessment (FRAIA) tool developed by the Dutch Ministry of Interior and Kingdom Relations.[436]

434 European Union Agency for Fundamental Rights (2022). Bias in Algorithms – Artificial Intelligence and Discrimination, available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf
435 White House (2022). Blueprint for an AI Bill of Rights, available at: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/
436 Dutch Ministry of the Interior and Kingdom Relations (2022). Impact Assessment Fundamental rights and algorithms, available at: https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms.

Human rights impact assessments (HRIAs)

HRIAs can assist in the identification of risks that judicial operators might not otherwise foresee in AI development and deployment. To accomplish this, HRIAs prioritise human rights implications over the optimization of the technology or its outputs. HRIAs or comparable processes could assure respect for human rights by design throughout the technology's lifecycle.

HRIAs evaluate technology based on a wide variety of potential human rights impacts. When ADM is used in the judicial settings, stakeholders should conduct transparent, impartial, and inclusive HRIAs, which consist of an examination of the products, services, and systems of intermediaries surrounding AI development and deployment and their effects on human rights. These HRIAs must incorporate input from affected communities and stakeholder organizations, including civil society and marginalized groups. The results of HRIAs should be made public and should be freely accessible and comprehensible. [437]

HRIAs for AI must investigate the inner workings of algorithms, i.e., they must analyse their technical components. HRIAs for algorithms must also be undertaken across the entire life cycle of an AI system, beginning with the earliest stages of its conception and continuing through important phases of its development and deployment. They should not be ex ante or ex post endeavours only. The Fundamental Rights and Algorithm Impact Assessment (FRAIA), developed by the Dutch government, and the Human Rights, Ethical and Social Impact Assessment in AI, developed by Alessandro Manterelo at the University of Turin, are recent HRIAs that meet these requirements. Both HRIAs give recommendations to assist AI developers and deployers in identifying the impact of AI systems on a broad range of basic rights. In addition, they provide various instances of potential mitigating strategies to prevent adverse effects. All of this minimizes the likelihood of unjustifiable human rights violations. The FRAIA considers the impact of AI systems on more than a hundred fundamental rights and sub-rights — for example, freedom of expression is subdivided into numerous sub-rights, such as freedom of the press, academic freedom, and whistleblowing — and proposes a comprehensive list of preventive and mitigating measures to limit infringements on these rights.

Below is a snapshot of the FRAIA process:

This fundamental rights and algorithm impact assessment (‹FRAIA›) is a discussion and decision-making tool for government organisations. The tool facilitates an interdisciplinary dialogue by those responsible for the development and/or use of an algorithmic system. The commissioning client is primarily responsible for the (delegated) implementation of the FRAIA.

The FRAIA comprises a large number of questions about the topics that need to be discussed and to which an answer must be formulated in any instance where a government organisation considers developing, delegating the development of, buying, adjusting and/or using an algorithm (hereinafter use of AI) When an algorithm is used, the FRAIA may serve as a tool for reflection. The discussion about the various questions should take place in a multidisciplinary team consisting of people with a wide range of specialisations and backgrounds. Per question, the FRAIA indicates who should be involved in the discussion. This tool pays attention to all roles within a multidisciplinary team, which are included in the diagram the diagram below. However, the list is not exhaustive. Likewise, the role or function names may differ from one organisation to another.

Role	FRAIA Part 1	FRAIA Part 2	FRAIA Part 3	FRAIA Part 7
Interset group	•			
Managment	•			
Citizen panel	•			
CISO or CIO	•	•		
Communications specialist		•	•	
Data scientist		•	•	
Data controller or datasource owner		•		
Domain expert (Employee who has domani knowlege redg the algorithm›s scope of application)	•	•	•	•
Data protection officer		•		
HR staff member			•	
Legal advisor	•	•	•	•
Algorithm developer		•		
Commissioning client	•	•	•	
Other project team members	•			
Project leader	•	•	•	•
Strategic ethics consultant		•	•	

Source: OECD, AI in Society, available at: https://www.oecd.org/en/publications/artificial-intelligence-in-society_eedfee77-en.html; Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, available at: https://www.lawfaremedia.org/article/assessing-impacts-of-ai-on-human-rights-it-s-not-solely-about-privacy-and-nondiscrimination

437 OSCE (2022). Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual, available at: https://www.osce.org/representative-on-freedom-of-media/510332

The Human Rights, Democracy, and the Rule of Law Assurance Framework (HUDERAF) for AI systems

The HUDERAF, proposed by the Alan Turing institute (which has been advising the CAHAI – the Council of Europe Ad hoc Committee on Artificial Intelligence) aims to present a coherent and integrated method for assessing the potential negative effects on human rights, democracy, and the rule of law caused using AI systems, as well as for ensuring that identified risks posed by AI to judicial operators are adequately mitigated and managed. The framework is made up specifically of several well-articulated but logically connected procedures and tools. It combines transparent risk management, impact mitigation, and innovation assurance approaches with context-based risk assessments and appropriate stakeholder involvement. Judicial operators could use the HUDERAF framework in assessing the potential negative impacts of AI on human rights.

The HUDERAF encompasses four components:

(1) The Preliminary Context-Based Risk Analysis (PCRA) gives a first indication of the context-based risks an AI system may pose to human rights, democracy, and the rule of law. The PCRA's major goal is to assist AI project teams in developing a reasonable strategy for risk management and assurance procedures as well as the degree of stakeholder involvement required throughout the project lifecycle.

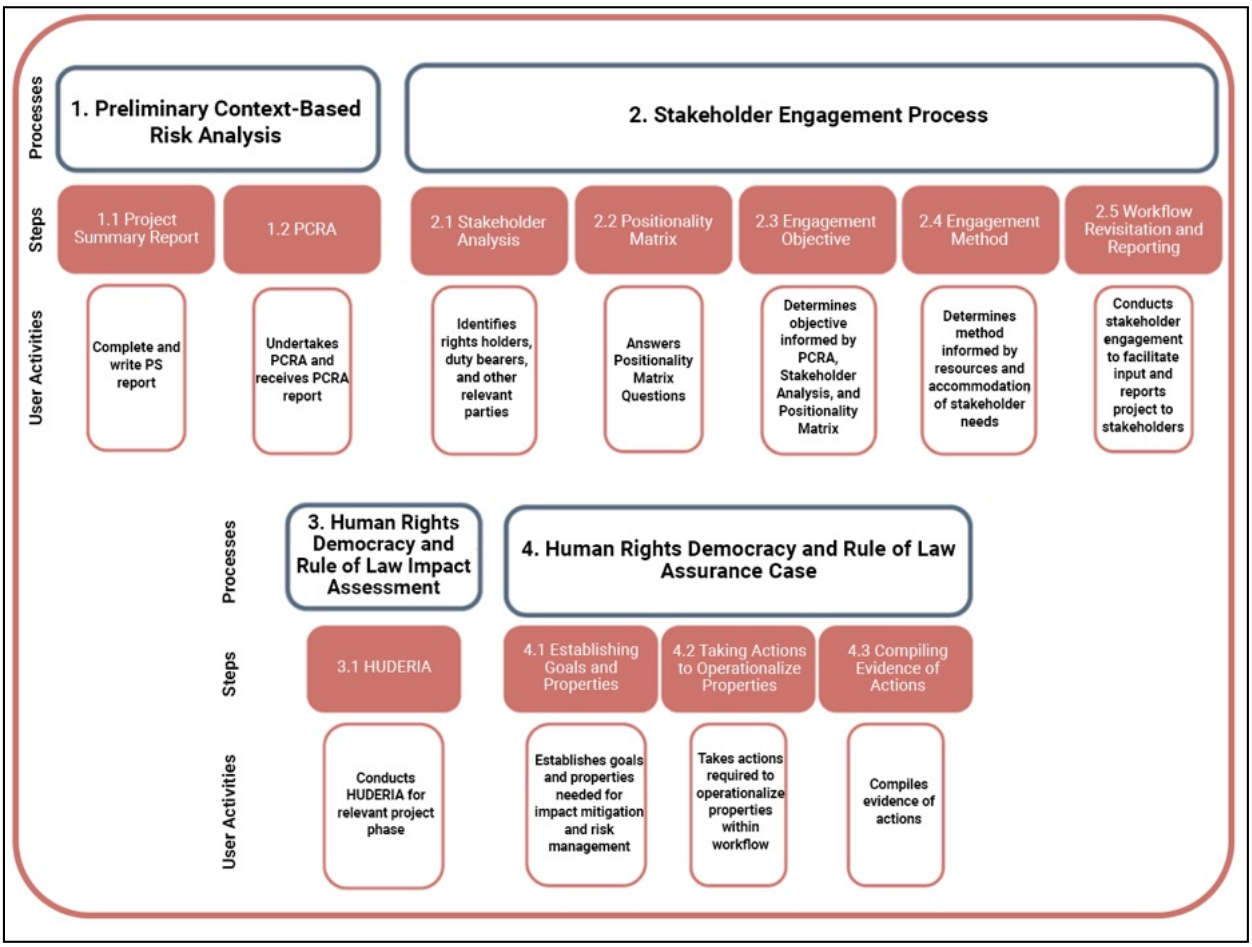
(2) The Stakeholder Engagement Process (SEP) supports appropriate stakeholder involvement and input throughout the project process by assisting project teams in identifying stakeholder salience. Through stakeholder participation, revisitation, and review, this method protects the equality and the contextual accuracy of HUDERAF governance processes.

(3) The Human Rights, Democracy, and the Rule of Law Impact Assessment (HUDERIA) gives project teams and involved stakeholders the chance to work together to create in-depth assessments of the possible and actual effects that the design, development, and use of an AI system could have on human rights, democracy, and the rule of law. Through the integration of stakeholder perspectives, this process contextualizes and validates previously identified potential harms, allows for the discovery of additional harms, allows for the collaborative assessment of the seriousness of identified potential adverse impacts, facilitates the co-design of an impact mitigation plan, establishes access to remedy, and establishes protocols for impact monitoring and re-assessment.

(4) The Human Rights, Democracy, and Rule of Law Assurance Case (HUDERAC) enables AI project teams to construct a structured justification that gives stakeholders demonstrable assurance that claims about the accomplishment of objectives set forth in the HUDERIA and other HUDERAF governance processes are justified in light of the evidence

at hand. Creating an assurance case facilitates internal reflection and discussion, encouraging the adoption of best practices and incorporating them into the design, development, and deployment lifecycles. Additionally, it offers a clear means to notify impacted stakeholders of the steps taken throughout the project workflow to reduce risks and guarantee the identification of pertinent normative objectives. A carefully constructed assurance case provides a transparent and easily understood framework for managing risks and mitigating their effects, supporting the right levels of social acceptance, accountability, and openness.[438]

Figure 14: Human Rights Democracy, and the rule of Law Assurance Framework (HUDERAF)



Source: Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer. The Council of Europe, available at: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf

438 Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, available at: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf

These group activities are intended to encourage the training participants to discuss and debate instances of possible human rights encroachments using ADM and AI in judicial operations and instances of judicial deliberation of human rights infringed using AI by third parties.

Activity 1

Please review Appendix B of the Canadian Directive on Automated Decision-Making and examine the four levels of impact that a decision assisted by AI can have on fundamental rights.[439]

Consider the following scenario: The Employment Agency in Country X intends to calculate the probability of registered job seekers finding employment within a certain period in the future, taking into account several factors: job seekers' age group, gender, education, health conditions, caring duties, the performance of their regional labour market and how long they have been registered with Employment Agency. Based on the calculated probability, job seekers will be assigned into different groups: group one that covers job seekers with high market opportunities, another group with medium and a last group with low opportunities. The AI system will assist the Employment Agencies' counsellors in assessing job seekers' opportunities and enable a more efficient use of resources. Based on this scenario, the training participants examine the four levels of impact that a decision made by the AI system can have on the rights of the job seekers.[440]

Appendix B: Impact Assessment Levels	
Level	Description
I	The decision will likely have little to no impact on: • the rights of individuals or communities, • the health or well-being of individuals or communities, • the economic interests of individuals, entities, or communities, • the ongoing sustainability of an ecosystem. Level I decisions will often lead to impacts that are reversible and brief.
II	The decision will likely have moderate impacts on: • the rights of individuals or communities, • the health or well-being of individuals or communities, • the economic interests of individuals, entities, or communities, • the ongoing sustainability of an ecosystem. Level II decisions will often lead to impacts that are likely reversible and short-term.
III	The decision will likely have high impacts on: • the rights of individuals or communities, • the health or well-being of individuals or communities, • the economic interests of individuals, entities, or communities, • the ongoing sustainability of an ecosystem. Level III decisions will often lead to impacts that can be difficult to reverse and are ongoing. At least level III would be probably reached for predictive policing activities in consideration of the high impact on the freedoms and rights of individuals and communities previously highlighted.

439 See: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

440 Barros Vale S., Zanfir-Fortuna G. (2022). Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities, available at: https://fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/

Appendix B: Impact Assessment Levels	
IV	The decision will likely have very high impacts on: • the rights of individuals or communities, • the health or well-being of individuals or communities, • the economic interests of individuals, entities, or communities, • the ongoing sustainability of an ecosystem. Level IV decisions will often lead to impacts that are irreversible and are perpetual.

Activity 2

Risk assessment factsheet template – look at the following risk assessment template and see if you can think of any other questions that you might ask to evaluate the risk assessment tool.

- Who created the risk assessment? Are they a public or private organization?
- How large was the training dataset?
- How was the training data set collected and assembled (i.e., what jurisdiction(s) is it from)?
- Over what time frame was the data collected?
- What factors (i.e., defendant characteristics) were included in the dataset? This question pertains to all the factors that were available about defendants, not necessarily all the factors that were used to train or develop the model.
- Does the dataset include instances of defendants who were detained? If so, does the data include outcomes for those people (i.e., did the data account for counterfactual estimation; if so, how)?
- Are there any known issues or errors with the data?
- In what year was the risk assessment created?
- What factors, among all the factors in the training data, were considered in the development of the risk assessment? If not, all factors were considered, how were those that were considered chosen?
- How were factors that were considered ultimately chosen for exclusion or inclusion in the final model (the risk assessment itself)?
- Does the final model include as a factor(s) arrest that did not lead to convictions? Does the final model include socioeconomic factors such as housing and employment status? Does the final model include personal health factors such as mental health or substance abuse? [split up into multiple questions if relevant info is available]
- How were weights assigned to each factor included in the final model? (Rounding correlation coefficients, Burgess Method, etc.)
- How does the final model define outcomes (i.e., during the model development process, was there a distinct outcome defined for each type of failure (failure to appear, new crime, new violent crime, etc.) or were outcomes compounded?

- What does the output of the model look like (i.e., a score on a scale of 1-10, etc.)?
- Does the model output risk level designations or convert raw scores into risk level designations such as "low risk," "moderate risk," and "high risk"?
- What proportion of samples in the training data set failed at each risk score and/or level (for example, what percentage of people with a score of 5 or a label of "moderate risk" actually failed to appear)?
- Did the model developers assess the predictive validity of the model? If so, how?
- Where is the risk assessment used?
- Are the factors and weights of the risk assessment publicly available?
- Does the risk assessment cost money for a jurisdiction to adopt?
- Does the adoption of the risk assessment require training? If so, by who?
- Does the risk assessment come with any sort of software or software package?
- Does the risk assessment involve or require an in-person interview?
- How does the risk assessment account for missing information?
- Has the risk assessment been analyzed on non-training data for predictive validity? Has the risk assessment been analyzed with training data or non-training data about performance for different race groups? Has the risk assessment been analyzed with training data or non-training data about performance for different genders? If so, by who, when, and using what data?[441]

Activity 3

Please discuss the following questions with other training participants:

- What does privacy entail in an era where real-time data collection is commonplace and there is a chance of data breaches, identity theft, or online fraud?
- Can we express ourselves freely on all digital tools and platforms without worrying about AI censorship or profiling?
- Can everyone have equal access to reliable information given the widespread dissemination of damaging material and lies online?
- How can we ensure that AI technologies help close the digital divide rather than widening already-existing disparities?

441 See: https://law.stanford.edu/pretrial-risk-assessment-tools-factsheet-project/

5. Resources

1. Access Now (2018). Mapping artificial intelligence strategies in Europe: a new report by Access Now, available at: https://www.accessnow.org/mapping-artificial-intelligence-strategies-in-europe/

2. Article 19, The Danish Institute for Human Rights (2017). Sample ccTLD Human Rights Impact Assessment Tool, available at: https://www.article19.org/wp-content/uploads/2017/12/Sample-ccTLD-HRIA-Dec-2017.pdf

3. Australian Human Rights Commission (2018). Final Report: Human Rights and Technology, available at: https://humanrights.gov.au/our-work/technology-and-human-rights/publications/final-report-human-rights-and-technology

4. CAHAI (2020). Legal Framework for AI Systems. Feasibility study of a legal framework for the development, design and application of artificial intelligence, based on Council of Europe's standards on human rights, democracy and the rule of law, Council of Europe Study, DGI/2021/04, available at: https://edoc.coe.int/en/artificial-intelligence/9648-a-legal-framework-for-ai-systems.html

5. Council of Europe (2020). Recommendation CM/Rec (2020) of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, available at: https://rm.coe.int/09000016809e1154

6. Dearden L. (2018). New Technology Can Detect ISIS Videos before They Are Uploaded, available at: http://www.independent.co.uk/news/uk/home-news/isis-videos-artificialintelligence-propaganda-ai-home-office-islamic-state-radicalisation-asi-data-a8207246.html

7. Duarte N., Llanso E., Loup A. (2017), Mixed Messages? The Limits of Automated Social Media Content Analysis, available at: https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/

8. Elsayed-Ali S. (2017). Artificial Intelligence and the Future of Human Rights, available at: https://medium.com/amnesty-insights/artificial-intelligence-and-the-future-of-human-rights-b58996964df5

9. Edwards L. (2022). Expert opinion: Regulating AI in Europe. Four problems and four solutions, available at: https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/

10. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act), available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

11. European Commission (2021). Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe, available at: https://artificialintelligenceact.eu/wp-content/uploads/2022/06/AIA-COM-Impact-Assessment-3-21-April.pdf

12. European Data Protection Supervisor. Necessity & Proportionality, available at: https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en

13. ICO, AI and data protection risk toolkit, available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/

14. Jones K. (2023), available at: https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights

15. Latonero M. (2018). Artificial Intelligence & Human Rights: A Workshop at Data & Society, available at: https://points.datasociety.net/artificial-intelligence-human-rights-a-workshop-at-data-society-fd6358d72149

16. Latonero M. (2019). Governing Artificial Intelligence: Upholding Human Rights and Human Dignity, available at: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

17. L.iberty (2020). Liberty wins ground-breaking victory against facial recognition tech, available at: https://www.libertyhumanrights.org.uk/issue/liberty-wins-ground-breaking-victory-against-facial-recognition-tech/

18. Mounk Y. (2018). Verboten. Germany's risky law for stopping hate speech on Facebook and Twitter, available at: https://newrepublic.com/article/147364/verboten-germany-law-stopping-hate-speech-facebook-twitter

19. OECD AI Policy Observatory. Live data, available at: https://oecd.ai/en/data?selectedArea=ai-research&selectedVisualization=top-countries-in-ai-scientificpublications-in-time-from-scopus

20. OECD AI Policy Observatory. Principles Overview, available at: https://oecd.ai/en/ai-principles

21. Ortiz Freuler J., Iglesias C. (2018). Algorithms and Artificial Intelligence in Latin America: A Study of Implementation by Governments in Argentina and Uruguay, World Wide Web Foundation.

22. Peralta Gutiérrez (2022). Marco normativo de la Inteligencia Artificial en el ámbito comparado. In: Herrera Triguero F., Peralta Gutiérrez A., Torres López L.S., El derecho y la inteligencia artificial, EUG: Granada 189–222.

23. Pielemeier J. (2018). The Advantages and Limitations of Applying the International Human Rights Framework to Artificial Intelligence, available at: https://medium.com/datasociety-points/the-advantages-and-limitations-of-applying-the-international-human-rights-framework-to-artificial-291a2dfe1d8a

24. Reiling D., Contini F. (2022). E-Justice Platforms: Challenges for Judicial Governance, International Journal for Court Administration, 13 (1), available at: https://iacajournal.org/articles/10.36745/ijca.445

25. Reisman D., Schultz J., Crawford K., Whittaker M. (2018). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, available at: https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2

26. Reitman R. (2022). Podcast Episode: Algorithms for a Just Future, available at: https://www.eff.org/deeplinks/2022/01/podcast-episode-algorithms-just-future

27. Stankovich M. (2021). Regulating AI and Big Data Deployment in Healthcare: Proposing Robust and Sustainable Solutions for Developing Countries' Governments, available at: https://www.dai.com/uploads/regulating-ai-cda.pdf

28. Vincent J. (2019). AI won't relieve the misery of Facebook's human moderators, available at: https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms

29. YouTubeHelp. How Content ID Works, available at: https://support.google.com/youtube/answer/2797370?hl=en

30. UN report on the rights to freedom of peaceful assembly and of association in the digital age A/41/41, 17 May 2019, available at: https://undocs.org/A/HRC/41/41

31. Taxonomy of Human Rights Risks Connected to Generative AI, United Nations Human Rights Office of the High Commissioner, available at: https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf

32. AI in Business and Finance, OECD Business and Finance Outlook, 2021, p.72, available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/09/oecd-business-and-finance-outlook-2021_377c2c18/ba682899-en.pdf

33. Ashraf, C. (2020) Artificial intelligence and the rights to assembly and association, Journal of Cyber Policy, 5:2, 163-179, DOI: 10.1080/23738871.2020.1778760

34. Gaumond E., Régis C., (2023), Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, Lawfare, available at: https://www.lawfaremedia.org/article/assessing-impacts-of-ai-on-human-rights-it-s-not-solely-about-privacy-and-nondiscrimination

35. Leloup D., Reynaud F., (2023) Social media: The cornerstone of manipulation and disinformation operations, Le Monde, available at: https://www.lemonde.fr/en/pixels/article/2023/02/16/social-media-the-cornerstone-of-manipulation-and-disinformation-operations_6016031_13.html#

36. Li S., Ham J., Eastin M., (2024), Social media users' affective, attitudinal, and behavioral responses to virtual human emotions, Telematics and Informatics, Volume 87, available at: https://doi.org/10.1016/j.tele.2023.102084

37. UN report on the rights to freedom of peaceful assembly and of association in the digital age A/41/41, 17 May 2019, available at: https://undocs.org/A/HRC/41/41

38. Taxonomy of Human Rights Risks Connected to Generative AI, United Nations Human Rights Office of the High Commissioner, available at: https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf

39. AI in Business and Finance, OECD Business and Finance Outlook, (2021), p.72, available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/09/oecd-business-and-finance-outlook-2021_377c2c18/ba682899-en.pdf

40. Cameran Ashraf (2020) Artificial intelligence and the rights to assembly and association, Journal of Cyber Policy, 5:2, 163-179, DOI: 10.1080/23738871.2020.1778760

41. Gaumond E., Régis C., (2023) Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, Friday, Lawfare, available at: https://www.lawfaremedia.org/article/assessing-impacts-of-ai-on-human-rights-it-s-not-solely-about-privacy-and-nondiscrimination

Suggested UNESCO Resources

UNESCO materials

Publications

Global Toolkit for Judicial Actors: International Legal Standards on Freedom of Expression, Access to Information and Safety of Journalists

- ☐ Available in: Arabic, Chinese, English, French, Portuguese, Russian, and Spanish

Toolkit for the Judiciary in Africa on the Legal Standards on Freedom of Expression

- ☐ Available in: English; French; and Portuguese

Guidelines for prosecutors on cases of crimes against journalists

- ☐ Available in: Amharic; Arabic; Chinese; Dari; English; French; Indonesian; Italian; Khmer; Portuguese; Pashto; Russian; Somali; Spanish; Swahili; Thai; Ukrainian; and Uzbek

Guidelines for judicial actors on privacy and data protection

- ☐ Available in: Arabic, Chinese, English, French, Portuguese, Russian; and Spanish

COVID-19: Guidelines on the role of judicial operators in the protection and promotion of the right to freedom of expression

- ☐ Available in: Arabic; Burmese; Chinese; English; French; Khmer; Portuguese; Russian; and Spanish

Safety of journalists covering protests: preserving freedom of the press during times of turmoil

- ☐ Available in: Arabic; Burmese; Chinese; English; French; Portuguese; Russian; and Spanish

Global toolkit for law enforcement agents: freedom of expression, access to Information and safety of journalists

- ☐ Available in: Arabic; English; French; Spanish; Chinese; Portuguese and Russian

Brochure on Freedom of expression and public order: fostering the relationship between security forces and journalists

- Available in: English; Portuguese; Russian; Ukrainian; and Somali

The "misuse" of the judicial system to attack freedom of expression: trends, challenges and responses

- Available in: Arabic, Chinese, English, French, Italian, Portuguese, Russian and Spanish

UNESCO guide for amicus curiae interventions in freedom of expression cases

- Available in: Arabic, Chinese, English, French, Russian and Spanish

Videos and webinar series

The Next Frontier: Intellectual Property in the Era of Generative Artificial Intelligence

- Available in: English and Spanish

The Admissibility Challenge: AI-Generated Evidence in the Courtroom

- Available in: English

Internet Governance Forum 2021 — Artificial Intelligence and the Rule of Law in the Digital Ecosystem

- Available in: English

Internet Governance Forum 2022 — Why Digital Transformation and Artificial Intelligence Matter for Justice

- Available in: English

UNESCO Video Explainers — How to stop impunity for crimes against journalists

- Available in: Arabic; Chinese; English; French; Russian; and Spanish

The Three-part Test: legitimate limits to freedom of expression

- Available in: Arabic; Chinese; English; French; Portuguese; Russian; and Spanish

The Rabat Plan of Action on the Prohibition of Incitement to Hatred: legitimate limits to freedom of expression

- Available in: Arabic; Chinese; English; French; Portuguese; Russian; and Spanish

UNESCO Video Explainers — What would a world without independent media look like?

- Available in: Arabic; Chinese; English; French; Russian; and Spanish

UNESCO Video Explainers — Why #FreedomOfExpression and #AccessToInformation are so central for free and fair elections?

- Available in: Arabic; Chinese; English; French; Russian; and Spanish

UNESCO Video Explainers — Regional Judicial Courts in Africa and Landmark Jurisprudence on Freedom of Expression

- Available in: English; French; and Portuguese

Legal challenges related to freedom of expression amid the COVID-19 pandemic

- Available in: English, French and Spanish

Courses

Massive Open Online Course (MOOC) on Artificial Intelligence and the Rule of Law

- Available in: Arabic, Chinese, English, French, Portuguese, Russian and Spanish

New Bonavero Institute-UNESCO multilingual Massive Open Online Course (MOOC) on freedom of expression and safety of journalists

- Available in: English, Arabic, Chinese, French, Portuguese, Russian and Spanish

How to use this toolkit?

This toolkit encourages an experiential pedagogical model: it is not intended to be prescriptive, and users are encouraged to draw on their own experiences considering the relevant contexts in which the toolkit is being used. Although it is aimed primarily at judicial operators, it may similarly be of use to a variety of others, including civil society organizations. There are several different ways in which the toolkit can be used as a resource:

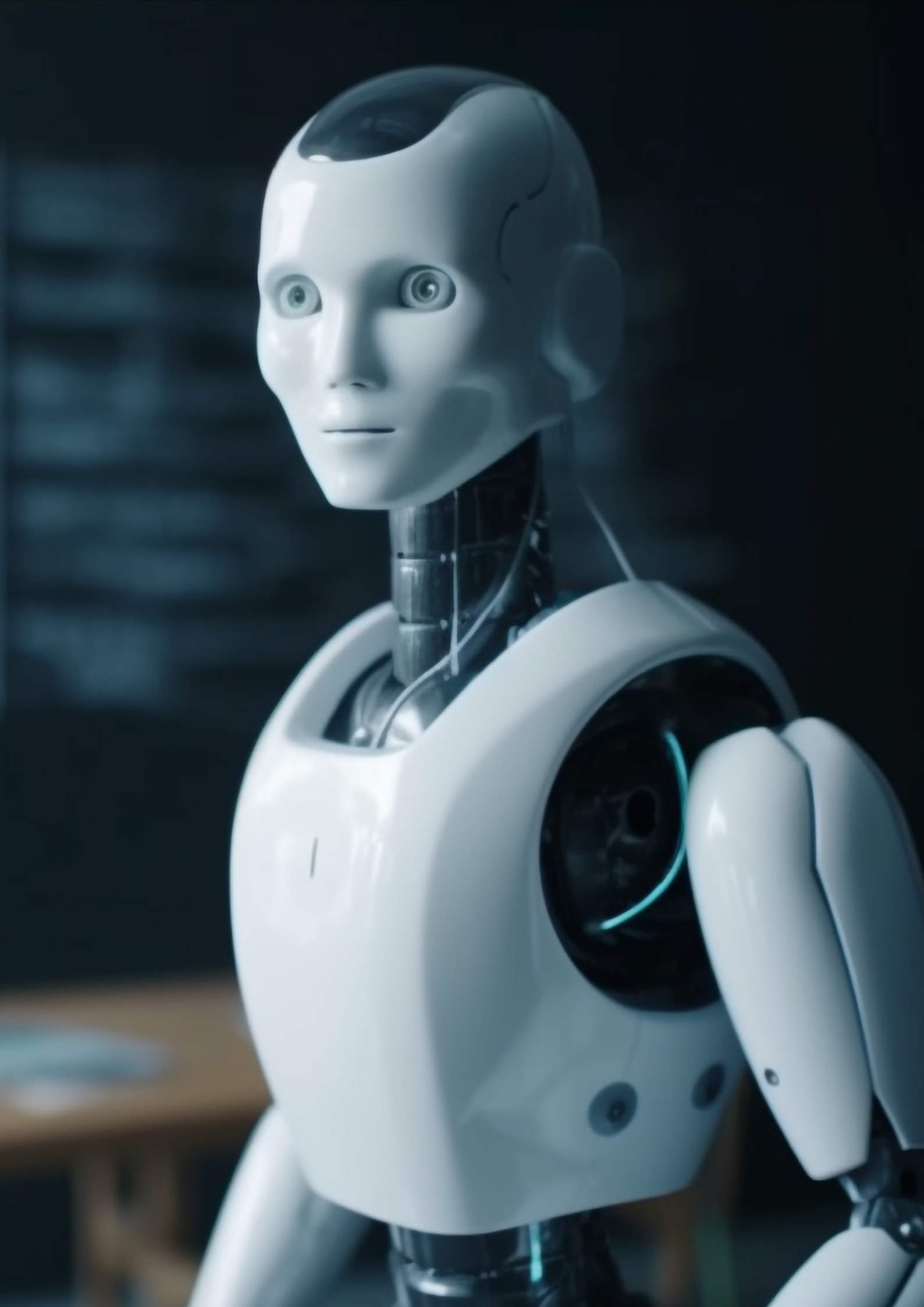
- Comprehensive in-person workshop: We would advise that a comprehensive in-person workshop covering all four modules should be at least three days. In circumstances where participants are not familiar with the fundamental principles of international human rights law, we would advise that the workshop take place over at least four days.
- Targeted workshop: Workshops could also be held on selected modules within the toolkit. In such circumstances, trainers should still ensure that the foundation is laid from the other modules that may be necessary for the participants to fully grasp the concepts and complete the tasks.
- Combined online course (such as a massive open online course) and in-person workshop: This format would provide more time for participants to engage with the materials and the self-assessment exercises, before being brought together in the in-person setting. Ideally, the online component should be supported with online discussion forums and other support.
- Self-study: The toolkit is self-explanatory in nature and can serve as a useful self-study resource to be engaged in individually or amongst a group of individuals working in a particular organisation. While there is often benefit in there being collaborative discussions and sharing of experiences, it can also be a useful starting point and reference for someone seeking to increase their understanding of emerging issues in AI and human rights.

Although longer workshops will allow for more activities to be engaged in, it is unlikely that there will be time to conduct all the suggested activities. This is at the discretion of the trainers. Trainers should seek to gauge from the groups the aspects that are of most relevance to the participants and can best be integrated into their work and domestic scenarios.

It is advised that trainers circulate a questionnaire to participants beforehand to ascertain their experience in this area of the law. The following template could be adapted depending on the expected training participants:

Global Toolkit for Judicial Operators on AI and the Rule of Law	
Participant details Name: Organisation: Designation: Country	Participant experience Do you have legal experience? Do you have experience in automated decision making, AI and human rights? Please explain.
Which modules would be of most use to you and your work? Please select. ☐ Module 1. Introduction to AI and the Rule of Law ☐ Module 2. AI adoption in the Judiciary ☐ Module 3. Legal and ethical challenges of AI deployment in the Judiciary ☐ Module 4. Human Rights and AI: Governance, regulation, and policy	
Please explain.	What are your objectives for this training?

"AI and human rights" is a dynamic and evolving area of the law. As such, there are likely to be frequent new developments. Instructors should be cognisant to stay abreast of these developments and update the training material accordingly.



Annex I

UNESCO Ethical Impact Assessment for AI systems

This instrument has two goals: First, to assess whether specific algorithms are aligned with the values, principles and guidance set up by the Recommendation. And second, to ensure transparency by calling for information about AI systems and the way they were developed to be available to the public. This is not how it works today, even for basic information about AI safety and reliability.

Impact Assessment tools are gaining ground to assess the true impact of AI systems. In fact, impact assessments are mandated by the EU AI Act for high-risk systems, and they are proposed as part of the Council of Europe's discussion on a Convention for AI.

The UNESCO Recommendation is unique in that it considers the entire AI lifecycle. The Ethical Impact Assessment therefore includes ex-ante and ex-post requirements. At an early stage, it establishes the importance of ensuring quality and representativeness data, the diversity of the teams developing the products, the robustness and transparency of the algorithms, their auditability, and the possibility of inserting check points at different moments of the development process.

The EIA is proposed to procurers of AI systems, as this is one of the main channels in which algorithms make their way to highly sensitive public domains. But the questions and the structure of the document are designed so the tools can also be used more generally by developers of AI systems, in the public or private sectors, who wish to develop AI ethically and fully comply with international standards such as the Recommendation.

The document comprises two main parts that together strike a balance between procedure and substance. In the first part, related to scoping, the goal is to understand the basics of the system, as well as to lay out some preliminary questions, such as whether automation is the best solution for the case at hand. It also raises questions about the project team and whether plans are in place to engage different stakeholders. The second part is dedicated to implementing the principles in the UNESCO Recommendation.

For each principle, questions will aim to assess:

a. Whether sufficient procedural safeguards have been put in place to ensure the system complies with the Recommendation; and

b. The (potential) positive outcomes and adverse impacts that may arise from the procurement and deployment of the system, specific to its context of use.

The Assessment Tool is available at: https://unesdoc.unesco.org/ark:/48223/pf0000386276

Annex II

Examples of additional activities

- Case studies of AI systems in public services in Latin America - http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf
- Interactive activity (courtroom algorithm game) of using Compas - https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/
- Trustworthy AI Playbook - https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook-executive-summary.pdf
- Algorithmic impact assessment activity, Government of Canada, Algorithmic Impact Assessment. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
- Assessment List for Trustworthy AI (ALTAI) exercise

AI mapping tool[442]

#	Question	Answers
1	What is the name of the Artificial Intelligence tool being assessed with this questionnaire?	
2	Briefly describe the tool's main functionality.	
3	What is motivating the use of AI tools in this case? (Check all that apply)	1) Existing backlog of work or cases 2) Improve overall quality of decisions 3) Lower transaction costs of an existing program 4) The tool is performing tasks that humans could not accomplish in a reasonable period of time 5) Use innovative approaches 6) Other
4	How was this tool developed?	1) Completely developed by your institution's technical staff 2) Developed in collaboration with an external entity 3) Procured, developed entirely by an external party 4) I don't know 5) Other

442 Brehm K., Hirabayashi M., Langevin C., Munoscano B.R., Sekizawa K., Shu J. (2020). The future of ai in the brazilian judicial system - ai mapping, integration and governance. The Future of AI in the Brazilian judicial System. AI Mapping, Integration, and Governance, Technical report, ITS Rio, available at: https://itsrio.org/wp-content/uploads/2020/06/SIPA-Capstone-The-Future-of-AI-in-the-Brazilian-Judicial-System-1.pdf

#	Question	Answers
5	Which e-Justice platform is this tool developed for/with?	
6	What stage of development is the tool currently in?	1) In development / ongoing procurement process 2) Prototype / Testing 3) Ready for deployment, not currently operating 4) Fully deployed
7	Which methods is the tool based on?	1) Logistic regression 2) Support Vector Machines 3) Decision Trees / Random Forest 4) Neural Networks / CNN 5) Oversampling / Resampling methods 6) Dimensionality reduction methods (PCA, Clustering, Manifold Learning) 7) Other:
8	Please check which, if any, of the following capabilities apply to the tool. (Check all that apply)	1) Modelling & risk assessment: Analysing data sets to identify patterns and recommend courses of action and in some cases trigger specific actions. 2) Data organization: Analysing data to categorize, process, triage, personalize, and serve specific content for specific contexts. 3) Image and object recognition: Analysing data to automate the recognition, classification, and context associated with an image or object. 4) Text and speech analysis: Analysing data to recognize, process, and tag text, speech, voice, and make recommendations, classifications or other kind of outputs based on the tagging. 5) Process optimization & workflow automation: Analysing data to identify anomalies, cluster patterns, predict outcomes or ways to optimize; and automate specific workflows. 6) None / Non applicable 7) Other
9	Does the tool perform any kind of analysis of unstructured data?	1) Yes 2) No 3) I don't know.
10	Is the data that was used to train the tool known by the team using it?	1) Yes 2) No 3) I don't know. 4) Not applicable

#	Question	Answers
11	Is the tool's code publicly available and reviewable?	1) Yes 2) No 3) I don't know. 4) Not applicable
12	Is the tool's algorithm and its code?	1) Open source 2) Court Owned 3) Owned by a third party
13	Is the tool collecting and/or analysing personal data (as defined by the General Data Protection Law)?	1) Collecting 2) Analysing 3) Neither
14	Is the tool collecting and/or analysing personally identifiable information?	1) Collecting 2) Analysing 3) Neither
15	The data used by the tool... (Check all that apply)	1) Was collected by a court, or a government entity. 2) Is publicly available and reviewable 3) Is shared with another entity. 4) Was collected by an external entity. 5) Is shared with an external entity.
16	Can the technical staff in your institution explain:	1) What the inputs of the tool are. 2) What the outputs of the tool are. 3) The process through which the inputs become outputs.
17	Can non-technical staff in your institution explain:	1) What the inputs of the tool are. 2) What the outputs of the tool are. 3) The process through which the inputs become outputs.
18	Has the tool gone through:	1) A technical monitoring and quality assurance processes 2) A review of its training data to detect biases 3) A legal and/or administrative review 4) Other

Annex III

Training agenda – template

Global Training Toolkit on AI and the Rule of Law for the Judiciary
3 Day Training

Date:

Title	Training for the[insert target audience] on the UNESCO Global Training Toolkit on AI and the Rule of Law for the Judiciary
Modality	Physical
Target Audience	
Dates	
Duration	3 days
Description	The training programme is based on the Global Toolkit on AI and the Rule of Law for the Judiciary
Organization	The training will be organized by the UNESCO
Registration deadline	
Training fees	
Language	

1. LEARNING OBJECTIVES

This training programme is intended to provide judicial operators with access to information and tools necessary to understand and consider the benefits of Artificial Intelligence ("AI") for their operations. At the same time, the training programme will help the Judiciary recognize AI's drawbacks and risks, including bias, discrimination, black boxes, and lack of accountability and transparency. The training programme will help judicial operators make better judgments and reduce potential human rights risks by offering guidance and perspectives on the principles, regulations, and relevant case law that underpin the use of AI responsibly in judicial contexts, and in general.

To balance the opportunities and challenges that AI technologies can present for the justice sector, the UNESCO Recommendation on the Ethics of AI highlights that "Member States should enhance the capacity of the Judiciary to make decisions related to AI systems as per the rule of law…". Hence the importance of this training programme for elaborating how the justice sector can take advantage of AI technologies and ensure that they are used ethically, responsibly, and in accordance with the international human rights law framework.

2. LEARNING OUTCOMES

After completing the training programme, judicial operators will be able to:

- Gain understanding of AI and Algorithmic Decision Making (ADM) and its use in the judicial processes and operations.
- Understand that AI is not neutral, and it is a socio technical system that represents the world around us.
- Build an ability to examine legal cases related to the use of AI.
- Understand the key issues related to algorithmic bias (such as gender bias, racial bias, intersecting forms of bias, etc.) and black boxes and explain why these are important in judicial settings.
- Get acquainted with the most recent regulatory measures and case law related to algorithmic bias, inappropriate use of algorithms in decision-making, including in contravention of the law, and black boxes.
- Understand and explain AI's impact on the following fundamental rights: privacy, freedom of expression, access to information, protection against discrimination, right to access to court, fair and impartial trials and hearings, and due process of law.

3.TARGET AUDIENCE

The training's primary target audience consists of judicial operators, primarily focusing on judges. The training can also include prosecutors, state attorneys, public lawyers, other justice sector stakeholders worldwide, and legal technology companies.

4. ENTRY REQUIREMENTS

Reading: Global Toolkit on AI and the Rule of Law for the Judiciary

5. INSTRUCTORS

NAME OF INSTRUCTORS	CONTACT DETAILS

6. TRAINING COURSE CONTENTS

The training programme is primarily based on the Global Toolkit on AI and the Rule of Law for the Judiciary and will cover the following topics:

1. Module 1: Introduction to AI and the rule of law
2. Module 2: AI adoption in the Judiciary
3. Module 3: Legal and ethical challenges of AI deployment
4. Module 4: Human rights and AI

7. TRAINING SCHEDULE CONTENTS AND AGENDA

Day 1: Introduction to AI and its use in the Judiciary

Time	Agenda
8:30 – 9:00	Participant sign-in and registration
9:00 – 9:30	Opening and Introduction to the training programme objectives
9:30 – 11:00	Session 1: Understanding AI and its building blocks Facilitator: This session aims to provide a comprehensive understanding of AI by exploring its definition and key building blocks. Through engaging discussions, illustrative examples, case studies and group activities we will examine the various components of AI systems, including algorithms, machine learning, data, and computational models. This session will also touch upon the key risks related to AI development and deployment, such as bias, black boxes and cybersecurity. By the end of this session, participants will gain a solid grasp of the key concepts related to AI, enabling them to navigate the field with confidence and clarity.
11:00 – 11:30	Coffee break

11:30 – 13:00	Session 2: What are the uses of AI in the justice sector? Facilitator: This session will outline some of the key uses of AI in the Judiciary, such as e-discovery and document review, the use of generative AI to assist with drafting documents, predictive analytics and ADM support, risk assessment tools, dispute resolution, language recognition and analytics, digital file, and case management.
13:00 – 14:30	Lunch
14:30 – 16:00	Session 3: Case studies on AI use in the Judiciary Facilitator: This session will examine select case studies on AI deployment in the justice system, such as VICTOR, Brazil, Singapore's Intelligent Court Transcription System, Prometea, Argentina, PretorIA, Colombia, Use of AI in China's justice system, Use of AI in India's justice system, UK's HART (Harm Assessment Risk Tool), PredPol, and Palantir. The session will invite participants to share their experience with AI systems and facilitate a broader conversation on the opportunities, challenges and risks associated with using these systems in the Judiciary.
16:00 – 16:30	Feedback and Assessment
16:30 – 16:45	Conclusion of day one and outline of the agenda for day two

Day 2: Legal and ethical issues related to AI systems

Time	Agenda
8:30 – 9:00	Participant sign-in and registration
9:00 – 11:00	Session 4: Algorithmic accountability and transparency Facilitator: Through insightful discussions and real-world case studies, this session will lead the participants through the concepts of algorithmic transparency and accountability concepts and highlights the key legal issues that judicial operators need to be aware of. Special focus will be given to biometric identification, facial recognition, and deepfakes.,
11:00 – 11:30	Coffee break

11:30 – 13:00	Session 5: Emerging case law on bias and black boxes Facilitator: The session will present existing case law that deals with algorithmic black boxes and bias in ADMs and AI systems used in public service delivery and by private sector. Using real-life case studies participants will discuss how bias and black boxes have resulted in infringement of human rights or any other harm, and how courts in different jurisdictions have dealt with this. Participants will discuss the issues of liability for harm done by these systems, as well as the use of AI for evidentiary purposes. The cases discussed will include: Deliveroo Case (2021), Foodinho Case (2021), People v. Chubbs (2015), State of New Jersey v. Francisco Arteaga, State v. Loomis, People v. Alvin Davis, State of New Jersey v. Pickett, Uber case concerning the use of the fraud-detection programme Mastermind United States v. Ellis, and the Australian case of Robodebt.
13:00 – 14:30	Lunch
14:30 – 16:00	Session 6: Ethical Impact Assessment of AI Systems Facilitator: This session will introduce the participants to key issues related to AI ethics, as well as key AI ethics frameworks on international, regional, and national levels. Using UNESCO's Ethical Impact Assessment of AI systems, the participants will assess hypothetical scenarios as part of breakout groups.
16:00 – 16:30	Feedback and Assessment
16:30 – 16:45	Conclusion of day one and outline of the agenda for day two

Day Three: AI and Human Rights

Time	Agenda
8:30 – 9:00	Participant sign-in and registration
9:00 – 11:00	Session 7: Human Rights and AI: right to access to court, fair trial, and due process, effective remedy, and rights to protection against discrimination. Facilitator: applications of AI have the potential to directly affect the equality of access to fundamental rights, including the right to privacy and the protection of personal information, the right to access to justice and the right to a fair trial, particularly regarding the presumption of innocence and the burden of proof, the right to employment, education, housing, and health, as well as the right to public services and welfare. If not accompanied by adequate safeguards against bias, AI technologies might contribute to denying access to rights disproportionately affecting women, minorities, and those who are already the most vulnerable and marginalized.
11:00 – 11:30	Coffee break

11:30 – 13:00	• Session 8: Human Rights and AI: (i) freedom of expression, (ii) right to privacy and data protection, and (ii) access to information. Facilitator: This session will present and discuss some of the human rights impacted by AI systems deployed by third parties and adjudicated by courts, such as freedom of expression, right to privacy and data protection, and access to information. The session will also discuss the relevant case law related to human rights and AI applications.
13:00 – 14:30	Lunch
14:30 – 16:00	Session 9: Emerging Issues at the Intersection of AI and Law Facilitator: - The session will briefly discuss concerns around: - Cybersecurity - Intellectual property rights - AI- generated evidence in courts - Use of AR and VR in courts
16:00 – 16:30	Feedback and Assessment
16:30 – 17:00	Summary and conclusion of the programme

8. METHODOLOGY (Didactic approach)

The training programme is based on the Global Toolkit on AI and the Rule of Law for the Judiciary. The Toolkit includes activities, content, and resources pertinent to AI, Human Rights, and the Rule of Law for judicial operators.

This training will be delivered physically, and include lectures, interactive exercises, and discussions. The training will be delivered using PowerPoint slides, selected reference materials, and daily self-assessment quizzes. The participants must revise, study, participate in scheduled activities and undertake self-assessments.

9. EVALUATION AND GRADING

Participants' performance in this training will be determined using a combination of grades for the participation sessions discussions and self-assessment quizzes.

- Participation in the sessions will be awarded 30 per cent.
- Self-assessments quizzes will be worth 70 per cent of the final grade of the training. There will be 6 questions per quiz.

In the end, Participants will receive a certificate of completion

10. TRAINING PRESENTATIONS

A number of presentations for each module and in different languages that can be used for the trainings are available at: Digital Innovation & Transformation (CI/DIT)

GLOBAL TOOLKIT ON

AI AND THE RULE OF LAW FOR THE JUDICIARY

Designed for judges, prosecutors, lawyers, judicial training institutions, and legal educators, the Global Toolkit on AI & the Rule of Law for the Judiciary is a practical gateway into the fast-evolving world of artificial intelligence in justice. It explains AI in clear, accessible terms, showing how algorithms are already shaping investigations, evidence assessment, case management, and even judicial decision-making. Through modular lessons, real-world examples, and guided reflections, it helps readers recognize both the promise of AI for efficiency and access to justice, and the dangers it poses to equality, transparency, privacy, and due process if left unchecked.

The Toolkit can be used directly in courses and workshops, inspiring new activities, teaching methods, and discussions that facilitate and improve the learning experience in judicial training institutions. It equips justice actors to ask the right questions, spot risks like bias and opacity, and uphold human rights standards whenever AI enters the courtroom.

Through this Toolkit and a growing portfolio of trainings, partnerships, and policy support, the AI & Rule of Law programme is empowering judiciaries worldwide to steer AI in service of justice.

unesco

9 789231 008795