UNITED ARAB EMIRATES
MINISTER OF STATE FOR ARTIFICIAL INTELLIGENCE,
DIGITAL ECONOMY & REMOTE WORK APPLICATIONS OFFICE

AI ETHICS

PRINCIPLES & GUIDELINES

DEC 2022

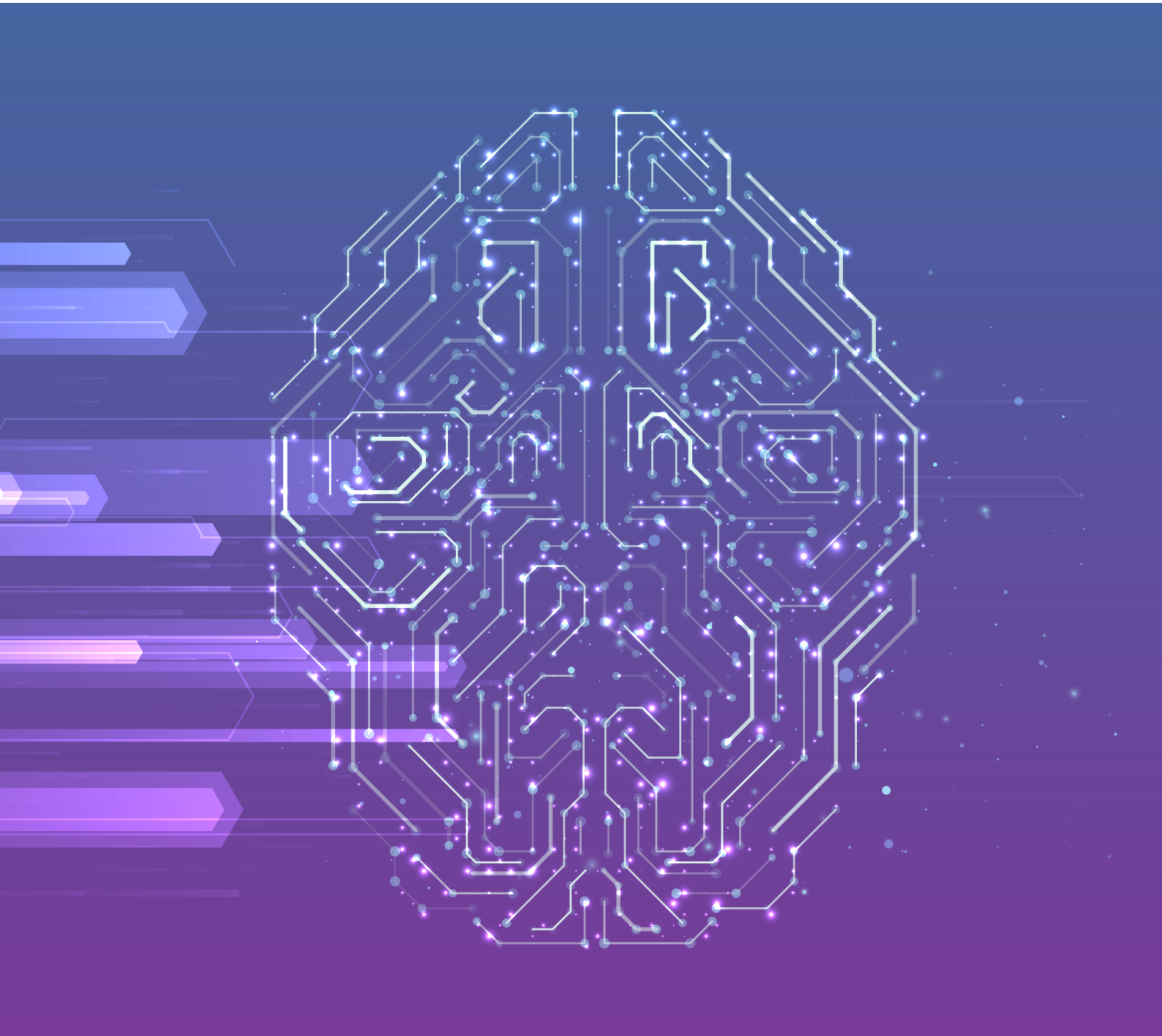
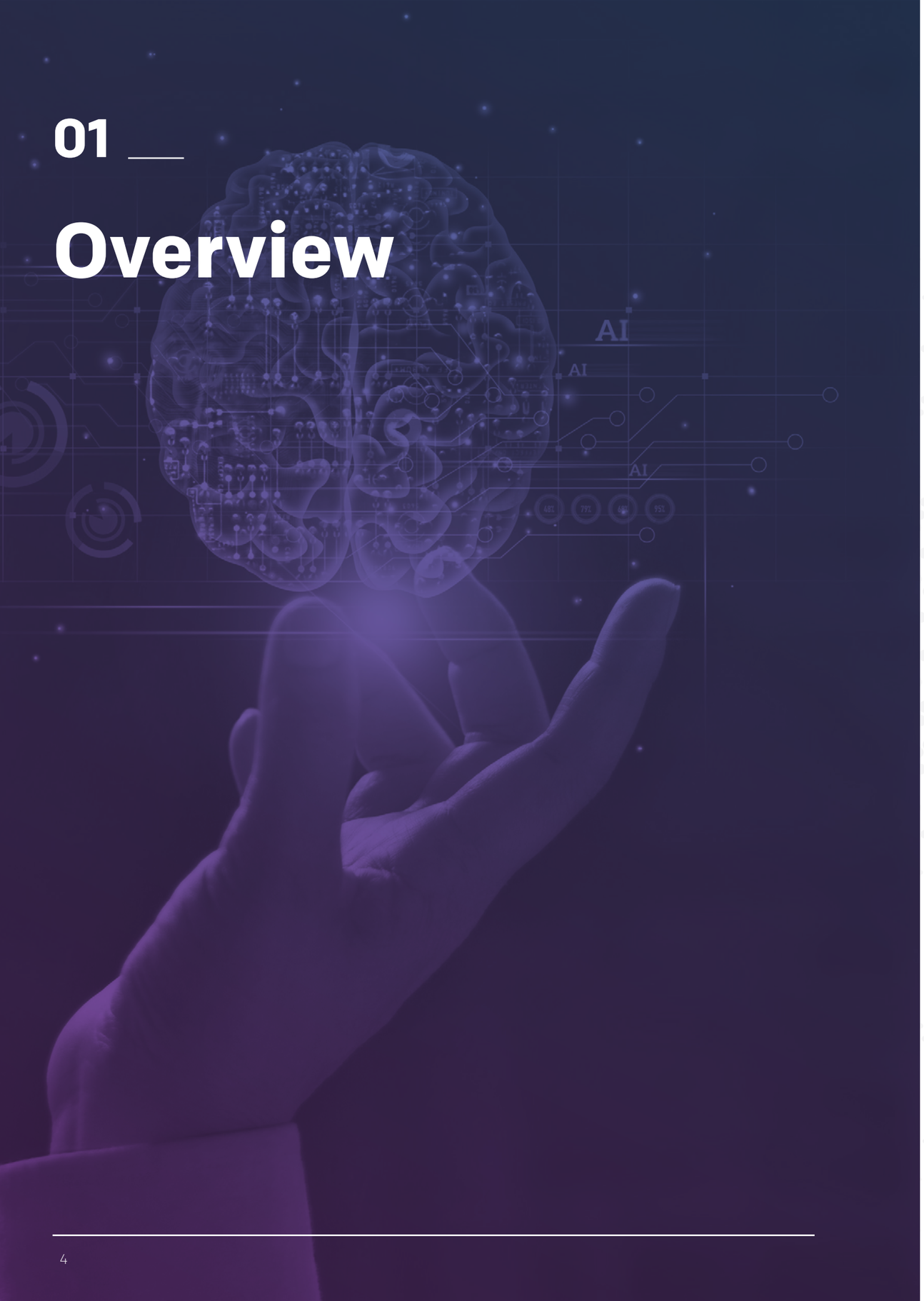


Table of Content

01 ___

Overview



1.1 Introduction

AI's rapid advancement and innovation potential across a range of fields is incredibly exciting. Yet a thorough and open discussion around AI ethics, and the principles organizations using this technology must consider, is needed.

This document - AI Ethics: Principles and Guidelines – intends to meet this need to balance these two central considerations of AI. It is designed to offer detailed guidance to help AI actors adhere to eight principles of AI ethics.

We will make AI systems that are:

explainable	**transparent**	**accountable**	**fair**
privacy preserving	**sustainable and environmentally friendly**	**human centered**	**robust, safe and secure**

The guidelines are non-binding, and are drafted as a collaborative multi-stakeholder effort, with full awareness of organizations' needs to innovate and protect their intellectual property. This is a collaborative process in which all stakeholders are invited to be part of an ongoing dialogue. We would like to see the AI Ethics Guidelines evolve into a universal, practical and applicable framework informing ethical requirements for AI design and use. The eventual goal is to reach widespread agreement and adoption of commonly agreed policies to inform the ethical use of AI nationally and around the world.

1.2 Scope

This document gives non-mandatory guidelines for achieving the ethical design and deployment of AI systems in both the public and private sectors. AI already surrounds us, but some applications are more visible and sensitive than others.

This document is applicable only to those AI systems which make or inform 'significant decisions' – that is, those decisions which have the potential for significant impact either on individuals or on society as a whole. They also apply to 'critical decisions', which are a subset of significant decisions and are of especially critical nature.

This document guides the further developments of sector specific requirements and principles, hence every entity within its sector can further amend, use, or complement its current frameworks to properly suit its context and the need of affected stakeholders.

It is a living document and will undergo further future reviews and enhancements.

1.3 Strategic alignment

This document comes as a fulfillment of the goals and aspirations set in the UAE National AI Strategy and the international Sustainable Development Goals (SDGs). The UAE has long been at the forefront of sustainable development. These shared global milestones have created global movement and adoption in several disciplines and sectors. As an early adopter, the UAE has established a national committee on SDGs with a long-term realization of the UAE SDG agenda 2030. These goals are further engrained and reflected in the AI Ethics Principles under the topics of fairness, inclusiveness, equality, human benefit, and rights.

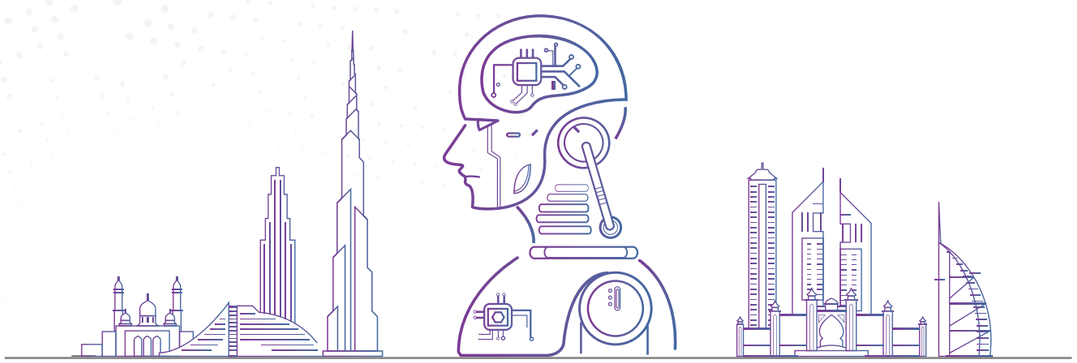
Additionally, the AI Office is currently developing additional policies to be taken into consideration along with this document:

AI Seal

The UAE AI Seal brand (UAI) will be used to attract talent and business from across the globe to come to the UAE to test and develop AI. This includes a UAI mark recognizing high quality, ethical AI companies. It would reward safe, efficient, verified AI technology with a 'UAI Seal of Approval'. The UAE AI Seal brand is currently under development by the AI Office.

AI Procurement Guidelines

The AI Procurement Guidelines will be used to guide Federal Government entities in the procurement of AI systems. They will list the general principles, proposed mechanisms for procurement and review of AI products between the federal government and vendors. The policy considers R&D for specific challenges that do not currently have an AI solution product that is market-ready. Based on global standards for AI Procurement, the policy aims to incentivize AI adoption within government entities and promote AI R&D within the private sector for specified government challenges. The procurement policy is currently under development by the AI Office.



1.4 Responsibility

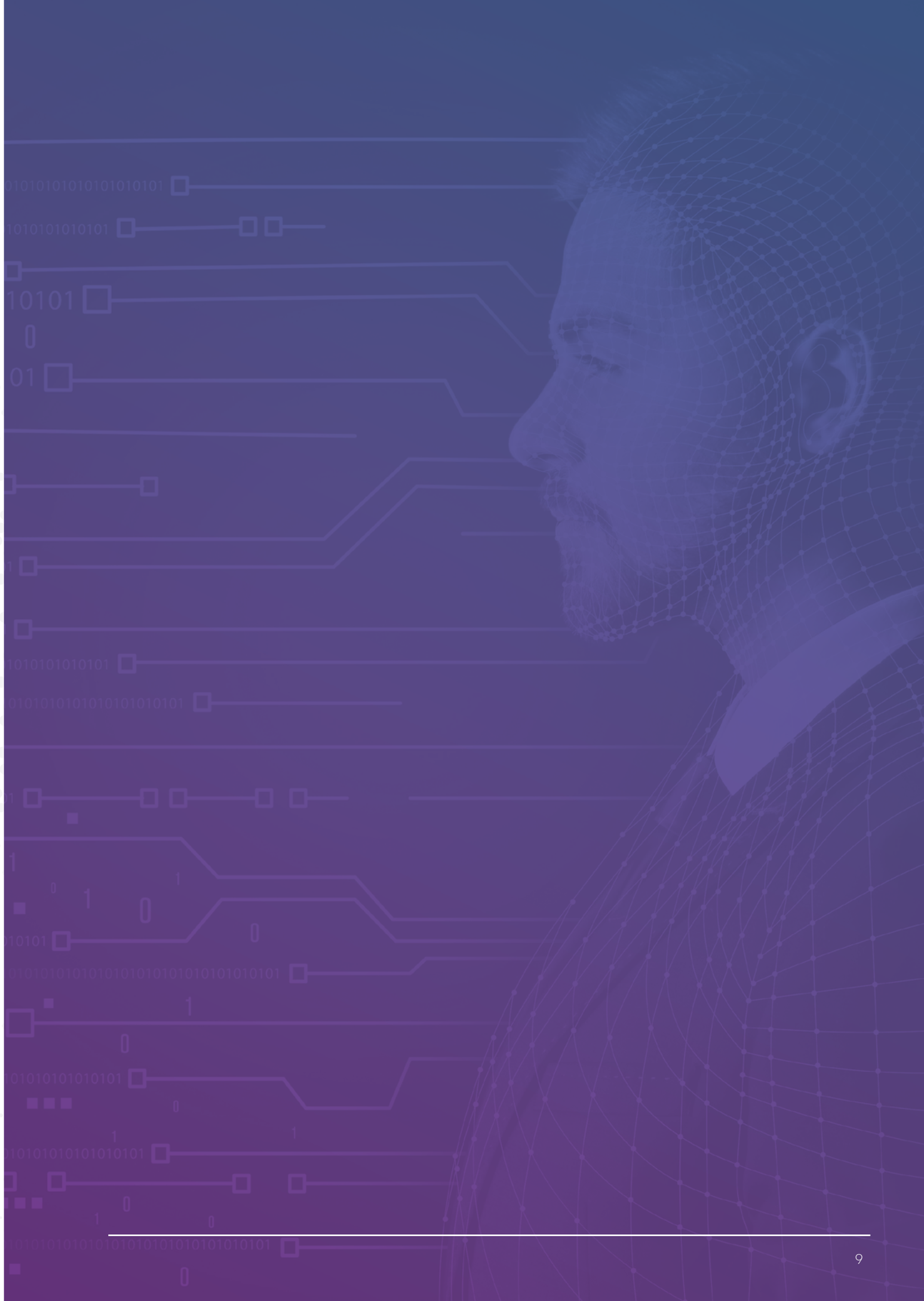
The AI Office will not be responsible for any misuse of the AI Ethics Principles and Guidelines. The user bears all the consequences of their use.

1.5 Licensing

This document is published under the terms of a Creative Commons Attribution 4.0 International License in order to facilitate its re-use by other governments and private sector organizations. In summary this means you are free to share and adapt the material, including for commercial purposes, provided that you give appropriate credit to the UAE AI and Blockchain Council as its owner and do not suggest the Minister of State for Artificial Intelligence, Digital Economy and Remote Work Applications Office endorses your use.

1.6 Toolkit

The document is part of a toolkit of self-governing nature that boosts awareness and enables government and private institutions to pursue innovative use cases while maintaining human values and principles.



02

Definitions

2.1 AI developer organization

An organization which does any of the following:

a. **determines the purpose of an AI system;**

b. **designs an AI system;**

c. **builds an AI system, or:**

d. **performs technical maintenance or tuning on an AI system**

Note:
The definition applies regardless of whether the organization is the ultimate user of the system, or whether they sell it on or give it away.

Example: A company develops an artificially intelligent facial recognition system and sells it to a country's border force, who use it to identify suspicious personnel. The company is an AI developer organization and the border force is an AI operator organization.

2.2 AI operator organization

An organization which does any of the following:

a. **uses AI systems in operations, backroom processes or decision-making;**

b. **uses an AI system to provide a service to an AI subject;**

c. **is a business owner of an AI system;**

d. **procures and treats data for use in an AI system; or**

e. **evaluates the use case for an AI system and decides whether to proceed.**

Notes:
1. This definition applies regardless of whether the AI system was developed in-house or procured.
2. It is possible for organizations to be both an AI developer organization and an AI operator organization.

2.3 Artificial Intelligence (also AI)

The capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning, learning and self-improvement[1].

2.4 Artificially Intelligent System (also AI system)

A product, service, process or decision-making methodology whose operation or outcome is materially influenced by artificially intelligent functional units.

Notes:

1. It is not necessary for a system's outcome to be solely determined by artificially intelligent functional units in order for the system to be defined as an artificially intelligent system; and
2. A particular feature of AI systems is that they learn behavior and rules not explicitly programmed in.

Example 1: A small claims court uses an artificially intelligent software package to collect evidence pertaining to a case, compare it to similar cases in the past, and present a recommended decision to a judge. The judge determines the final outcome. This decision-making methodology is materially influenced by an artificially intelligent functional unit, and is therefore classified as an AI system.

Example 2: A government entity uses a chatbot which allows customers to ask routine questions, book appointments and conduct minor financial transactions. The chatbot responds to customer queries with pre-written responses and is based on pre-programmed decision rules. Therefore, the chatbot is not an AI system. If, however, the chatbot autonomously adjusted its treatment of customers based on the outcome of past cases, it would be an AI system.

2.5 Bias (of a system)

Inclination or prejudice for or against one person or group, especially in a way considered to be unfair[2].

[1] Consistent with ISO/IEC 2382:2015

2.6 Critical decision

An individually significant decision which is deemed to either have a very large impact on an individual or to have especially high stakes. These can be especially sensitive, have the potential to cause high loss or damage, be societally significant, or set an important precedent.

Notes:

The types of decisions referred to here are the same as those in the definition of significant-at-scale decisions, except in this case the effects are felt as a result of an individual decision rather than an aggregate of many decisions..

Example: A court determines whether a defendant is guilty of a criminal charge, with the punishment for guilt being a life sentence. This is a critical decision because it has a very large impact on the life of the defendant and also sets precedent for similar cases in the future.

2.7 Functional Unit

Entity of hardware or software, or both, capable of accomplishing a specified purpose[3] .

2.8 Individually significant decision

A decision which has the potential for significant impact on at least one individual's circumstances, behavior or choices, or has legal or similarly significant effects on him or her.

Example: A company decides to make an employee redundant. This is an individually significant decision because of its potential impact on the employee's financial situation.

[2] Oxford Dictionary 2018, Oxford University Press, viewed online 4th October 2018, <https://en.oxforddictionaries.com>

[3] From ISO/IEC 2382:2015

2.9 Non-operational bias (of a system)

Bias that is either:

a. **not a design feature; or**

b. **not important in achieving the stated purpose of the system.**

2.10 Set of significant-at-scale decisions

A set of decisions made by the same system or organization which, when taken in aggregate, have significant impact on society as a whole or groups within it.

Notes:

1. The decisions need not be individually significant in order to qualify, in aggregate, as a set of significant-at-scale decisions; and
2. Examples of areas which have a large impact on society and which include but are not limited to: the large-scale allocation of resources or opportunities amongst groups; the structure of government; the division of power between large entities or groups; the law, and its interpretation and enforcement; conflict and war; international relations, etc.

Example: An AI system is used by a website to determine which content to show users. This decision is not individually significant, since a user is not greatly affected by whether a particular piece of media is shown to them. However, if the website is popular then the AI system may be making a set of significant-at-scale decisions, because any biases in the system will affect a large number of users.

2.11 Significant decision

A decision which is either individually significant or is part of a set of significant-at-scale decisions.

2.12 Subject of an artificially intelligent system (also AI subject)

A natural person who is any of the following:

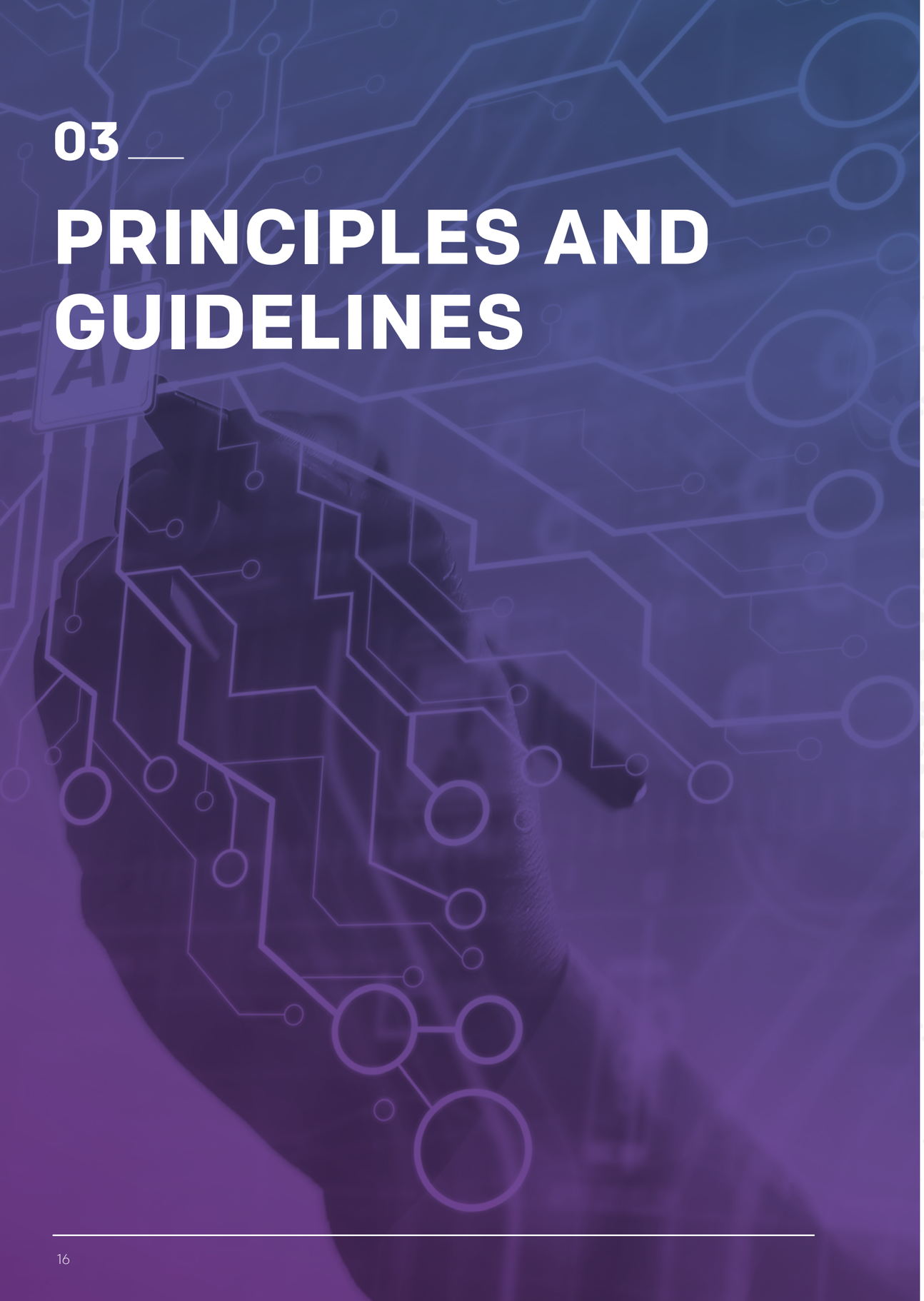
a. **an end-user of an AI system**

b. **directly affected by the operation of or outcomes of an AI system, or:**

c. **a recipient of a service or recommendation provided by an AI system**

03

PRINCIPLES AND GUIDELINES



3.1 Fairness (Principle 1)

3.1.1 Principle

We will make AI systems fair

- Data ingested should, where possible, be accurate and representative of the affected population.
- Algorithms should avoid non-operational bias.
- Steps should be taken to mitigate and disclose the biases inherent in datasets.
- Significant decisions should be provably fair.
- All personnel involved in the development, deployment and use of AI systems have a role and responsibility to operationalize AI fairness and should be educated accordingly.

3.1.2 Guidelines

3.1.2.1

Benefits of AI systems should be available and accessible to all. AI systems should not discriminate against people or groups in a way that could have an adverse impact. Especially significant and critical decisions made or assisted by AI should be provably fair.

a Considering that fairness can have many different definitions, an organization should document its own definition of fairness for the context that the AI system is going to be implemented in. Organizations should document what the implemented fairness objective stands for and why this choice was considered the most suitable for the given scenario.

b It is recommended to identify and document demographic groups that may be adversely impacted and mitigate the risk where possible.

c AI developers and AI operators could consider formal procedures such as Discrimination Impact Assessments as a means of ensuring fairness. This assessment should be documented as well as the mitigation measures that were implemented by the organization. The impact assessment should be conducted pre-release and regularly after as an ongoing evaluation, results are advised to be documented.

3.1.2.2
Consideration should be given to whether the data ingested is accurate and representative of the affected population

d Fairness has many different definitions in different cultures and for different contexts. Encouraging a diverse and inclusive AI ecosystem is thus all the more crucial to ensure that one definition of fairness does not contradict another, and that the process of defining fairness itself is fair, with under-represented groups represented in the discussion.

e AI developers and AI operators should undertake reasonable data exploration and/or testing to identify potentially prejudicial decision-making tendencies in AI systems arising from data inaccuracy and biases in the data.

f AI developers and operators should evaluate all datasets to assess inclusiveness of identified demographic groups and collect data to close any gaps.

g AI developers and AI operators should refrain from training AI systems on data that is not likely to be representative of the affected AI subjects, or is not likely to be accurate, whether that be due to age, omission, method of collection, or other factors.

Example: Following a natural disaster, a government relief agency uses an AI system to detect communities in greatest need by analyzing social media data from a range of websites. However, those communities where smartphone penetration is lower having less presence on social media, and so are at risk of receiving less attention. Therefore, the charity complements their AI tool with traditional techniques to identify needy populations elsewhere.

h AI developers should consider whether their AI systems can be expected to perform well when exposed to previously unseen data, especially when evaluating people who are not well-represented in the training data.

3.1.2.3
Consideration should be given to whether decision-making processes introduce bias

a When subjecting different groups to different decision-making processes, AI developers should consider whether this will lead to non-operational bias.

b When evaluating the fairness of an AI system, AI developers and AI operators should consider whether AI subjects in the same circumstances receive equal treatment.

Example: An organization uses an AI tool to automate the pre-screening of candidates for a job opening. It is trained on data from the company's existing employees, the majority of whom are from the same ethnic background. Therefore, the system learns to use name and nationality as discriminating factors in filtering job applicants. This could have been identified through testing and rectified by, for example, balancing the training data or only using relevant data fields for training.

3.1.2.4
Significant decisions informed by the use of AI should be fair

a AI developers and AI operators could consider formal procedures such as Discrimination Impact Assessments as a means of ensuring fairness.

3.1.2.4
AI operators should consider whether their AI systems are accessible and usable in a fair manner across user groups

3.1.2.5
Consideration should be given to the effect of diversity on the development and deployment processes

b Organizations should seek to include people from diverse demographic backgrounds across the full lifecyle, to include design, development and deployment processes. Organizations should seek to engage diverse internal and external groups also.

c AI developers should consider whether the assumptions they make about AI subjects could be wrong or are likely to lead to non-operational bias; if so, they should consider consulting the AI subjects in a representative manner during the design, development and deployment to confirm these assumptions.

3.2 Accountable AI (Principle 2)

3.2.1 Principle

We will make AI systems accountable

- Accountability for the outcomes of an AI system lies not with the system itself but is apportioned between those who design, develop and deploy it;
- Developers should make efforts to mitigate the risks inherent in the systems they design;
- AI systems should have built-in appeals procedures whereby users can challenge significant decisions;
- AI systems should be developed by diverse teams which include experts in the area in which the system will be deployed; and
- AI systems should be subject to external audit and decision quality assurance.

3.2.2 Guidelines

3.2.2.1
Accountability for the outcomes of an AI system should not lie with the system itself

a. Accountability for loss or damages resulting from the application of AI systems should not be attributed to the system itself.

b. AI operators and AI developers should consider designating individuals to be responsible for investigating and rectifying the cause of loss or damage arising from the deployment of AI systems.

[4] Cabinet Office (UK), Data Science Ethical Framework, Version 1.0, licensed under the Open Government License v3.0, p. 13

3.2.2.2
Positive efforts should be made to identify and mitigate any significant risks inherent in the AI systems designed

a. AI operators should only use AI systems that are backed by evidence-based academic or industrial research, and AI developers should base their development on such research.

Example: A foreign country has a government service which identifies parents who owe money in child maintenance. The data matching process is often incorrect due to misspelled names or missing data which results in some individuals being incorrectly targeted automatically by the system with the result being a large bill, poor credit ratings and even freezing wages. The recourse for individuals who are incorrectly targeted is time-consuming and not straightforward . If the potential impact of incorrect decisions had been assessed, mitigation measures such a user-friendly review procedure could have been set up.

b. AI operators should identify the likely impact of incorrect automated decisions on AI subjects and, in the case where incorrect decisions are likely to cause significant cost or inconvenience, consider mitigating measures.

c. AI operators should consider internal risk assessments or ethics frameworks as a means to facilitate the identification of risks and mitigating measures.

d. In designing AI systems to inform significant decisions, AI developers should consider measures to maintain data accuracy over time, including:

- the completeness of the data.
- timely update of the data, and.
- whether the context in which the data was collected affects its suitability for the intended use case.

Example: A border camera scanning for predictors of risk may misinterpret a "tic" of an individual with Tourette syndrome as suspicious. These can manifest in a diverse fashion, and should not cause this person to undergo secondary inspection every time they pass through the border . If the data is updated after the first case is encountered then it would avoid causing inconvenience on subsequent visits.

[5] Government of Canada. (2018). Responsible AI in the Government of Canada. Digital Disruption White Paper Series. Version 2.0, p.26. Retrieved from: https://docs.google.com/document/d/1Sn-qBZUXEUG4dVk909eSg5qvfbpNlRhzIefWPtBwbxY/edit
[6] Stoica, I. et. al.,2017, A Berkeley View of Systems Challenges for AI, p. 2, https://arxiv.org/pdf/1712.05855.pdf

e AI developers and AI operators should tune AI models periodically to cater for changes to data and/or models over time.

f AI operators should consider whether AI systems trained in a static environment will display model instability when deployed in dynamic environments.

Example: AI systems will need to be able to adapt to the changes in the environment that they are deployed in. For example, a self-driving car would need to quickly adapt to unexpected and dangerous road by learning in real time from other cars that have successfully dealt with these conditions. In addition, such mission-critical applications must handle noisy inputs and defend against malicious actors .

g AI developers should collaborate with AI Operators to train models using historical data from the Operator.

h AI Operators should consider working with their vendors (AI developers) to continually monitor performance.

i AI Operators should subject AI systems informing significant decisions to quality checks at least as stringent as those that would be required of a human being taking the same decision.

3.2.2.3
AI systems informing critical decisions should be subject to appropriate external audit

a When AI systems are used for critical decisions, external auditing of the AI systems in question should be used as a means to ensure meaningful standards of transparency and accountability are upheld.

b In the case that critical decisions are of civic interest, public release of the results of the audit should be considered as a means of ensuring public processes remain accountable to those affected by them.

c In the case that critical decision are life and death decisions, these decisions should be supported by further validation and verification via a human operator.

d Facilitate traceability and auditability of AI systems, particularly in critical contexts or situations.

e Be mindful that there might be fundamental tensions between different principles and requirements. Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

f Adopt a Trustworthy AI assessment list when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.

g Keep in mind that such an assessment list will never be exhaustive. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders in this.

3.2.2.4
AI subjects should be able to challenge significant automated decisions concerning them and, where appropriate, be able to opt out of such decisions

a AI operators using AI systems to inform significant decisions should provide procedures by which affected AI subjects can challenge a specific decision concerning them.

b AI operators should consider such procedures even for non-significant decisions.

c AI operators should make affected AI subjects aware of these procedures and should design them to be convenient and user-friendly.

d AI operators should consider employing human case evaluators to review any such challenges and, when appropriate, overturn the challenged decision.

Example: A bank allows customers to apply for a loan online by entering their data. The bank uses an AI system to automatically determine whether to give the loan and what the interest rate should be. They provide users with an option to contest the decision and have it reviewed by a human . They also require that customers justify their challenge by filling in a form, which assists the case reviewer and deters customers from challenging a decision without good reason.[7]

[7] Adapted from EU Commission, Can I be subject to automated individual decision-making, including profiling? Retrieved from: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rights-citizens/my-rights/can-i-be-subject-automated-individual-decision-making-including-profiling_en#example

e AI operators should consider instituting an opt-out mechanism for significant automated decisions.

f AI operators could consider "crowd challenge" mechanisms whereby a critical number of complaints triggers an investigation into the fairness and/or accuracy of a decision-making process as a whole.

3.2.2.5
AI systems informing significant decisions should not attempt to make value judgements on people's behalf without prior consent

a When informing an AI subject about significant choices they will make, AI systems should not unreasonably restrict the available options or otherwise attempt to influence their value judgements without the explicit consent of the AI subject in question.

3.2.2.6
AI systems informing significant decisions should be developed by diverse teams with appropriate backgrounds

a AI developers who develop AI systems which may be used to assist in making critical decisions should involve in the process experts with a background in social science, policy, or another subject which prepares them to evaluate the broader societal impact of their work.

b Development of AI systems informing significant decisions should include consultation with experts in the field in which the system will be deployed.

Example: An app that uses AI to assess medical symptoms and has a large user base had to face regulatory scrutiny because of number of complaints from doctors. They warned that the application can miss signs of serious illness. A number of different shortcomings were identified by doctors, some of which the company could address and resolve.[8]

3.2.2.7
AI operators should understand the AI systems they use sufficiently to assess their suitability for the use case and to ensure accountability and transparency

a In the case of critical decisions, AI operators should avoid using AI systems that cannot be subjected to meaningful standards of accountability and transparency.

b AI developers should consider notifying customers and AI operators of the use cases for which the system has been designed, and those for which it is not suitable.

[8] Financial Times. 2018. High-profile health app under scrutiny after doctor's complaints. Retrieved from: https://www.ft.com/content/19dc6b7e-8529-11e896-dd-fa565ec55929

3.3 Transparent AI (Principle 3)

3.3.1 Principle

We will make AI systems transparent

- **Developers should build systems whose failures can be traced and diagnosed.**
- **People should be told when significant decisions about them are being made by AI.**
- **Within the limits of privacy and the preservation of intellectual property, those who deploy AI systems should be transparent about the data and algorithms they use.**
- **Responsible disclosures should be provided in a timely manner, and provide reasonable justifications for AI systems outcomes. This includes information that helps people understand outcomes, like key factors used in decision making.**

3.3.2 Guidelines

3.3.2.1

Traceability should be considered for significant decisions, especially those that have the potential to result in loss, harm or damage.

a For AI systems which inform significant decisions, especially those with the potential to cause loss, harm or damage, AI developers should consider building in traceability (i.e. the ability to trace the key factors leading to any specific decision).

b Organizations should ensure that harms caused through AI systems are investigated and redressed, by enacting strong enforcement mechanisms and remedial actions, to make certain that human rights and the rule of law are respected in the digital world and in the physical world.

c To facilitate the above, AI developers and AI operators should consider documenting the following information during the design, development and deployment phases, and retaining this documentation for a length of time appropriate to the decision type or industry:

- the provenance of the training data, the methods of collection and treatment, how the data was moved, and measures taken to maintain its accuracy over time;
- the model design and algorithms employed; and
- changes to the codebase, and authorship of those changes.

d Where possible given the model design, AI developers should consider building in a means by which the "decision journey" of a specific outcome (i.e. the component decisions leading to it) can be logged.

Example: A technology company has a product which is designed to assist in medical diagnosis. It documents each stage of its reasoning and relates it back to the input data.[9]

[9] See IBM WatsonPaths

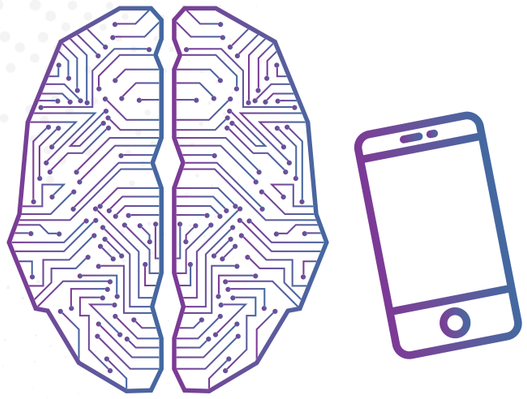
3.3.2.2
People should be informed of the extent of their interaction with AI systems

a AI operators should inform AI subjects when a significant decision affecting them has been made by an AI system.

Example: A small claims court adjudicates minor civil matters such as debt collection and evictions. They introduce an AI system to suggest the outcome of a ruling. At the time of the ruling the plaintiff and defendant are notified that the decision was assisted by an AI system. The court also provides an explanation for the decision.

b If an AI system can convincingly impersonate a human being, it should do so only after notifying the AI subject that it is an AI system.

Example: A technology company produces a conversational AI agent which can make some phone calls on behalf of its users. Those who receive the calls may believe that they are speaking to a human. Therefore, the company programs the agent to identify itself at the start of every conversation.



3.4 Explainable AI (Principle 4)

3.4.1 Principle

We will make AI systems as explainable as technically possible

- **Decisions and methodologies of AI systems which have a significant effect on individuals should be explainable to them, to the extent permitted by available technology.**
- **It should be possible to ascertain the key factors leading to any specific decision that could have a significant effect on an individual.**
- **In the above situation we will provide channels through which people can request such explanations.**

3.4.2 Guidelines

3.4.2.1
AI operators could consider providing affected AI subjects with a high-level explanation of how their AI system works

a AI operators could consider informing the affected AI subjects in understandable, non-technical language of:

- the data that is ingested by the system;
- the types of algorithms employed;
- the categories into which people can be placed; and
- the most important features driving the outcomes of decisions
- A comprehensive list of feature engineering and models that was considered during the model building phase.

Example: A person turned down for a credit card might be told that the algorithm took their credit history, age, and postcode into account, but not learn why their application was rejected[10] .

b For non-sensitive public sector use cases designed for the common good, AI operators could consider making source code, together with an explanation of the workings of the AI system, available either publicly or upon request (this should be done only if there is low risk of people 'gaming the system').

c The AI operator should maintain necessary documentation that provides elaboration and clarification on how the algorithm works, for example documentation of processes, decision making flow charts of the system, and how the appeal process is embedded.

d An individual should have the ability to contest and seek effective redress against decisions made by AI systems. These must be addressed by the group or team supporting these models.

e Develop appropriate impact indices for the evaluation of AI system technological interventions from multiple perspectives.

3.4.2.2
AI operators should consider providing affected AI subjects with a means to request explanations for specific significant decisions, to the extent possible given the state of present research and the choice of model

a AI operators should consider providing a means by which people affected by a significant decision informed by AI can access the reasoning behind that decision.

Example: The US Consumer Financial Protection Bureau requires that creditors in the US who reject credit applications must explain to the applicant the principal reason(s) why their application was rejected (e.g. "length of residence" or "age of automobile")[11] . In particular, "statements that the adverse action was based on the creditor`s internal standards or policies or that the applicant, joint applicant, or similar party failed to achieve a qualifying score on the creditor`s credit scoring system are insufficient".

b Where such explainability is not possible given available technology, AI operators should consider compromises such as counterfactual reasoning, or listing the most heavily weighted factors contributing to the decision.

[10] The Guardian. AI watchdog needed to regulate automated decision-making, say experts. Retrieved from: https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions

Example: The UK's NHS developed a tool called Predict, which allows women with breast cancer to compare their case to other women who have had the same condition in the past, and visualize the expected survival rate under various treatment options. The website has an explanation page which shows the weights behind various factors and contains a description of the underlying mathematics[12] .

c In the case that such explanations are available, they should be easily and quickly accessible, free of charge and user-friendly.

3.5 Robust, Safe and Secure AI (Principle 5)

3.5.1 Principles

AI systems will be technically robust

- **AI systems should be technically robust with a preventative approach to risks which operates in a manner such that they reliably behave.**
- **AI Developers should ensure AI systems will not cause any unintentional harm and adverse impacts.**
- **AI systems should be resilient to attacks and security such as data poisoning and model leakage.**
- **AI systems should have safeguards that enable a fallback plan in case of problems.**
- **AI system results should be reproducible.**

[11] Consumer Financial Protection Bureau, 12 CFR Part 1002 - Equal Credit Opportunity Act (Regulation B), § 1002.9 Notifications, Retrieved from https://www.consumerfinance.gov/policy-compliance/rulemaking/regulations/1002/
[12] Predict website, accessible at http://www.predict.nhs.uk/predict_v2.1/legal/algorithm

3.5.2 Guidelines

3.5.2.1

AI Operators should continue conducting vulnerability assessments, verification of AI system different behaviors in unexpected situations and for any dual-use case, to include:

a. Putting measures to ensure integrity and resilience of IA system against potential attacks.

Example: IBM researchers presented a novel approach to exploiting AI for malicious purposes. Their system, DeepLocker embeds AI capabilities within the malware itself in order to improve its evasion techniques. It uses an Artificial Intelligence model to identify its target using indicators like facial recognition, geolocation and voice recognition.

b. Verifying how AI systems behave in unexpected situations and environments.

c. Taking preventative measures against any possible system dual-case use.

3.5.2.2

AI operators should define a suitable fall back plan and test it to maintain readiness in unexpected situations and in high safety risks level, to include:

a. Developing a plan to measure and assess potential safety risks to you or any third party from technology use accidental or malicious misuse.

b. Define thresholds for system acceptable results and governance procure to fall back on alternative defined and tested plans.

c. Considering an insurance policy to mitigate risks arising from potential damages.

Example: Adopting a risk-based approach to procurement and clearly communicating it to vendors can help address such issues by giving the procuring organization advance notice of the specific oversight capabilities it will need in future stages of the system lifecycle, preventing vendors from presenting intellectual property arguments against required testing, monitoring and auditing of their AI systems going forward and rewarding vendors—of all sizes—that are more advanced and responsive in their responsible AI efforts.

3.5.2.3

AI Operators should assure end users of the system's reliability through documentation, and operationalizing processes for testing and verification of desired outcomes , specifically:

a. AI system should be designed with an approach to continue monitoring if the system meets goals, purposes and intended application.

b. AI Operators should define an approach to ensure results are reproducible through clear process and documentation.

c. AI Operators should publish documentation to assure system robustness to end users.

3.5.2.4

AI systems will be safe, secure and controllable by humans

a. Safety and security of people, be they operators, end-users or other parties, will be of paramount concern in the design of any AI system.

b. AI systems should be verifiably secure and controllable throughout their operational lifetime, to the extent permitted by technology.

c. The continued security and privacy of users should be considered when decommissioning AI systems.

d. AI systems that may directly impact people's lives in a significant way should receive commensurate care in their designs.

e. Such systems should be able to be overridden or their decisions reversed by designated individuals.

[13] IBM. https://documents.trendmicro.com/assets/white_papers/wp-malicious-uses-and-abuses-of-artificial-intelligence.pdf

3.5.2.5

Ai systems should not be able to autonomously hurt, destroy or deceive humans

a. AI systems should be built to serve and inform, and not to deceive and manipulate.

b. Nations should collaborate to avoid an arms race in lethal autonomous weapons, and such weapons should be tightly controlled.

c. Active cooperation should be pursued to avoid corner-cutting on safety standards.

d. Systems designed to inform significant decisions should do so impartially.

3.6 Human Centered AI (Principle 6)

3.6.1 Principles

We will give AI systems human values and make them beneficial to society

- **Government will support the research of the beneficial use of AI.**
- **Stakeholders throughout society should be involved in the development of AI and its governance.**
- **Design AI systems to adopt, learn and follow the norms and values of the community they serve.**
- **Systematic analyses that examine the ethics of designing affective systems to nudge human beings prior to deployment are needed.**
- **Decisions related to lethal force, life and death should not be delegated to AI systems. Rules and standards should be adopted to ensure effective human control over those decisions.**

3.6.2 Guidelines

3.6.2.1

AI should be developed to align with human values, contribute to human flourishing and benefit society.

a. Society should be consulted in a representative fashion to inform the development of AI.

b. Stakeholders throughout society should be involved in the development of AI and its governance.

c. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet.

d. Organizations should prioritize having all their stakeholders learn about wellbeing metrics as a potential determinant of how they create, deploy, market and monitor their AI technologies.

Example: IEEE P7010 recommended practice[14] establishes wellbeing metrics relating to human factors directly affected by intelligent and autonomous systems and establishes a baseline for the types of objective and subjective data these systems should analyze and include (in their programming and functioning) to proactively increase human wellbeing.

3.6.2.2

Systematic analyses that examine the ethics of designing affective systems to nudge human beings prior to deployment are needed.

a. Nudging in AI systems should have an opt-in system policy with explicit consent.

b. We recommend that when appropriate, an affective system that nudges human beings should have the ability to accurately distinguish between users, including detecting characteristics such as whether the user is an adult or a child. Additional protections must be put in place for vulnerable populations, such as children, when informed consent cannot be obtained, or when it may not be a sufficient safeguard.

c. AI systems with nudging strategies must be carefully evaluated, monitored, and controlled.

[14] IEEE. https://sagroups.ieee.org/7010/

3.6.2.3

Humanity should retain the power to govern itself and make the final decision, with AI in an assisting role.

a. Decisions related to lethal force, life and death should not be delegated to AI systems. Rules and standards should be adopted to ensure effective human control over those decisions.

b. Responsible parties (e.g., parents, nurse practitioners, social workers, and governments) should be trained to detect the influence due to AI and ineffective mitigation techniques. In the most extreme cases it should always be possible to shut down harmful AI system.

c. Those actions undertaken by an affective system that are most likely to generate an emotional response should be designed to be easily changed.

d. Users should be able to make informed autonomous decisions regarding AI systems. They should be given appropriate knowledge and tools to comprehend and interact with AI systems to satisfactory degree and, where possible to reasonably self-assess or challenge the system.

3.6.2.4

AI systems should be designed in a way that respects the rule of law, human rights, society values, and should include appropriate safeguards to ensure a fair and just society.

a. AI Developers should design AI systems to adopt, learn and follow the norms and values of the community they serve.

b. Organizations could identify the norms of the specific community in which AI systems are to be deployed in. In particular, pay attention to the norms relevant to the kinds of tasks that the AI systems are designed to perform. These could be documented as well as how these norms are addressed by the AI system.

c. To respond to the dynamic change of norms in society the AI system could be able to adjust its existing norms and learn new ones, while being transparent about these changes.

d. Designers should consider forms and metrics for assessing an AI system's norm conformity over the lifetime of the system (e.g. human-machine agreement on moral decisions, verifiability of AI decisions, justified trust).

e. The norm identification process must document the similarities and differences between the norms that humans apply to other humans and the norms they apply to AI systems. Norm implementations should be evaluated specifically against the norms that the community expects the AI system to follow.

f. In situations where it is needed, human rights impact assessments and human rights due diligence, human determination codes of ethical conduct, or quality labels and certifications intended to promote human centered values and fairness should be considered.

g. Mechanisms should be put into place to receive external feedback regarding AI systems that potentially infringe on fundamental rights.

3.7 Sustainable and Environmentally Friendly AI (Principle 7)

3.7.1 Principle

We will promote sustainable and environmentally friendly AI

- **AI systems should be used to benefit all human beings, including future generations. The environment is fundamental to this.**
- **Throughout the AI lifecycle, implementations should only be carried out on the basis of a full understanding and acknowledgement of implications for sustainability and environment.**
- **An AI system's development, deployment and use, should be assessed via critical examination of resource usage and energy consumption.**
- **Mechanisms to measure environmental impact due to type of energy use and processing power provided by data centers should be established.**

3.7.2 Guidelines

3.7.2.1 Throughout the AI lifecycle, implementations should always be carried out after full understanding and acknowledgement of AI implications on sustainability and environment.

a. The application of Artificial Intelligence to address sustainability challenges are well understood. Building and running green, sustainable AI systems is of increasing importance, given the large carbon footprint that they can generate and the wider context of addressing climate change.

b. AI model development, and by extension commissioning, should therefore seek to balance technical performance with energy consumption and environmental impact.

c. Efforts should be made to estimate and understand sustainability and environmental impact across the AI lifecycle (e.g. the energy consumption costs associated with model training, and CO2 emissions and cloud compute costs associated with the deployment and running of the system).

d. While there is no standardized means of reporting on the environmental impact of AI systems, where possible, model development and commissioning should make carbon impacts a core consideration alongside functional and business requirements.

e. Smaller models should be considered – shrinking down the model size and using fewer compute cycles (to balance financial and performance costs with the end performance of the model).

f. Carbon awareness should be considered – adjusting the operational parameters of the AI system to dynamically select the best time and location for energy use from the grid can reduce its carbon footprint.

Example: In agriculture, AI can transform production by better monitoring and managing environmental conditions and crop yields. AI can help reduce both fertilizer and water, all while improving crop yields.

Example: Google`s DeepMind division has developed AI that teaches itself to minimize the use of energy to cool Google`s data centers. As a result, Google reduced its data center energy requirements by 40% [15].

[15] https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40

3.8 Privacy Preserving AI (Principle 8)

3.8.1 Principle

We will respect people's privacy

- **AI systems should respect privacy and use the minimum intrusion necessary.**
- **AI systems should uphold high standards of data governance and security, protecting personal information.**
- **Surveillance or other AI-driven technologies should not be deployed to the extent of violating internationally and/or UAE's accepted standards of privacy and human dignity and people rights.**
- **Privacy by design should be embedded in AI systems and where possible AI algorithm should have adequate privacy impact assessments.**
- **Adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach and ensured throughout the life cycle of AI systems.**
- **AI developers and operators should strive to strike the balance between privacy requirements, individual rights and innovation growth and society benefits.**

3.8.2 Guidelines

3.8.2.1
Establish mechanism for users to flag issues related to privacy or data protection

a. AI Operators should review the system for proper consent logging, ability of users to revoke permission whenever applicable.

b. Consider training AI model without or with minimal user of potentially sensitive or personal data.

c. Use measures to enhance privacy such as encryption, anonymization and aggregation.

3.8.2.2
Establish an oversight mechanism for data collection, storage and processing and use across your organization

a. Assess who can access data and under which conditions.

b. Assign specific responsibilities and role for Data Protection Officers.

c. Prevention of harm to privacy necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

d. AI systems must guarantee privacy and data protection throughout a system's entire lifecycle. This includes the information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI systems generated for specific users or how users responded to particular recommendations). Digital records of human behavior may allow AI systems to infer not only individuals' preferences, but also their age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them.

e. In any given organization that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place in line with national privacy legislation (be these universal or sector-specific). These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual's data should be allowed to do so.

f Algorithmic systems require adequate privacy impact assessments, which also include societal and ethical considerations of their use and an innovative use of the privacy by design approach. AI actors need to ensure that they are accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system.

g Establish data policies or equivalent frameworks, or reinforce existing ones, to ensure full security for personal data and sensitive data, which, if disclosed, may cause exceptional damage, injury or hardship to individuals. Examples include data relating to offences, criminal proceedings and convictions, and related security measures; biometric, genetic and health data; and personal data such as that relating to race, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other characteristics.

h Promote mechanisms, such as open repositories for publicly funded or publicly held data, source code and data trusts, to support the safe, fair, legal and ethical sharing of data, among others.

i Promote and facilitate the use of quality and robust datasets for training, development and use of AI systems, and exercise vigilance in overseeing their collection and use.

Example: with the help of AI & ML algorithms[16] , Synthetic data can be created to enhance privacy. The data created will have the same statistical characteristics for the testing environment and where a third party can have access.

[16] https://research.aimultiple.com/privacy-enhancing-technologies/

04 ___

Bibliography

Microsoft. 2022. Microsoft Responsible AI Standard, v2. General Requirements. June 2022. Retrieved from:
https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4ZPmV

ISO. 2022. Proposed ISO Standard on Risk Management of AI: What Businesses Should Know. May 2022. Retrieved from:
https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi

GSMA 2022. The AI Ethics Playbook. Implementing ethical principles into everyday business. February 2022. Retrieved from:
https://www.gsma.com/betterfuture/wp-content/uploads/202201//The-Mobile-Industry-Ethics-Playbook_Feb-2022.pdf

OECD. 2022. AI Policy Observatory. Retrieved from:
https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

UNESCO. 2022. Recommendation on the Ethics of Artificial Intelligence. Adopted in November 2021. Retrieved from:
https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Retrieved from:
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Commission. High Level Expert Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy AI . April 2019. Retrieved from:
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Centre for Data Ethics and Innovation. UK Government: The roadmap to an effective AI assurance ecosystem. December 2021. Retrieved from:
https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem

The New York City Council. 2021. A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools. Retrieved from:
https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=%20B051915D-A9AC-451E-81F86596032-FA3F9&

UNESCO, Recommendation on the Ethics of Artificial Intelligence. 2021. Retrieved from:
https://unesdoc.unesco.org/ark:/48223/pf0000381137

AI NOW Institute. Algorithmic accountability for the public sector. August 2021. Retrieved from: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

IEEE 2019. Ethically Aligned Design. Version 2 – For Public Discussion. Retrieved from: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

IEEE 2017. Ethically Aligned Designed : A vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Retrieved from: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Google. AI Principles 2021. Retrieved from: https://ai.google/principles/

Microsoft. Responsible AI Standard v2 (June 2022). Retrieved from: https://www.microsoft.com/en-us/ai/responsible-ai

The ECONOMIST Group 2022 . Pushing forward: the future of AI in the Middle East and North Africa Report. Retrieved from: https://impact.economist.com/perspectives/sites/default/files/google_ai_mena_report.pdf

PDPC. (2018, June 5). Discussion paper on Artificial Intelligence (AI) and Personal Data. Singapore: Personal Data Protection Commission Singapore (PDPC). Retrieved from: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/Discussion-Paper-on-AI-and-PD---050618.pdf

ITI. AI Policy Principles. Retrieved from: https://www.itic.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf

Cabinet Office.(2016, May 19). Data Science Ethical Framework. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524298/Data_science_ethics_framework_v1.0_for_publication__1_.pdf

European Parliament. (2017, February 16). European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (20152103/(INL)). Retrieved from: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0++0051DOC+XML+V0//EN

Villani, C. (2018). For a meaningful Artificial Intelligence towards a French and European strategy. Retrieved from: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

CNIL. (2017). How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Retrieved from: https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

Executive Office of the President of the United States. National Science and Technology Council. Networking and Information Technology Research and Development Subcommittee. (2016). The national artificial intelligence research and development strategic plan. Retrieved from: https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

Executive Office of the President of the United States National Science and Technology Council. Committee on Technology. (2016). Preparing for the future of artificial intelligence. Retrieved from: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp NSTC/preparing_for_the_future_of_ai.pdf

The Headquarters for Japan's Economic Revitalization. (2015). New Robot Strategy. Japan's Robot Strategy. Vision, Strategy, Action Plan. Retrieved from: http://www.meti.go.jp/english/press/2015/pdf/0123_01b.pdf

Treasury Board of Canada Secretariat. (2018). Responsible Artificial Intelligence in the Government of Canada. Digital Disruption White Paper Series. Version 2.0. https://docs.google.com/document/d/1Sn-qBZUXEUG4dVk909eSg5qvfbpNlRhzIefWPtBwbxY/edit

The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems. Toronto, Canada: Amnesty International and Access Now. Retrieved from: https://www.accessnow.org/cms/assets/uploads/201808//The-Toronto-Declaration_ENG_08-2018.pdf

http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/oral/73546.html (house of lords select committee oral evidence, Q61)oral evidence, Q61)

House of Lords Select Committee on Artificial Intelligence. (2018). AI in the UK: ready, willing and able? https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100100/.pdf

House of Commons Science and Technology Select Committee. Algorithms in Decision Making. https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351351/.pdf

UNITED ARAB EMIRATES
MINISTER OF STATE FOR ARTIFICIAL INTELLIGENCE,
DIGITAL ECONOMY & REMOTE WORK APPLICATIONS OFFICE