SMART DUBAI

AI ETHICS PRINCIPLES & GUIDELINES

دبـــي الذكية
SMART DUBAI

دبـــي الذكية
SMART DUBAI

TABLE OF CONTENTS

دبـــي الذكية
SMART DUBAI

TABLE OF CONTENTS

دبـــي الذكية
SMART DUBAI

“OUR VISION IS FOR DUBAI TO EXCEL IN THE DEVELOPMENT AND USE OF AI IN WAYS THAT BOTH BOOST INNOVATION AND DELIVER HUMAN BENEFIT AND HAPPINESS.”

DR. AISHA BINT BUTTI BIN BISHR
Director General - Smart Dubai Office

RESPONSIBILITY

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

LICENSING

AI PRINCIPLES

دبـــي الذكية
SMART DUBAI

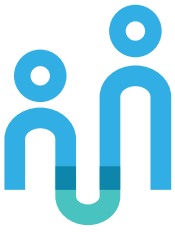
DUBAI'S AI PRINCIPLES

ETHICS

AI systems should be fair, transparent, accountable and understandable

SECURITY

AI systems should be safe and secure, and should serve and protect humanity



HUMANITY

AI should be beneficial to humans and aligned with human values, in both the long and short term

INCLUSIVENESS

AI should benefit all people in society, be governed globally, and respect dignity and people rights

AI PRINCIPLES

ETHICS

WE WILL MAKE AI SYSTEMS FAIR

- Data ingested should, where possible, be 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

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

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

AI PRINCIPLES

ETHICS

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

AI PRINCIPLES

SECURITY

AI SYSTEMS WILL BE SAFE, SECURE AND CONTROLLABLE BY HUMANS

- Safety and security of the people, be they operators, end-users or other parties, will be of paramount concern in the design of any AI system
- AI systems should be verifiably secure and controllable throughout their operational lifetime, to the extent permitted by technology
- The continued security and privacy of users should be considered when decommissioning AI systems
- AI systems that may directly impact people's lives in a significant way should receive commensurate care in their designs, and;
- Such systems should be able to be overridden or their decisions reversed by designated people

AI SYSTEMS SHOULD NOT BE ABLE TO AUTONOMOUSLY HURT, DESTROY OR DECEIVE HUMANS

- AI systems should be built to serve and inform, and not to deceive and manipulate
- Nations should collaborate to avoid an arms race in lethal autonomous weapons, and such weapons should be tightly controlled
- Active cooperation should be pursued to avoid corner-cutting on safety standards
- Systems designed to inform significant decisions should do so impartially

AI PRINCIPLES

HUMANITY

WE WILL GIVE AI SYSTEMS HUMAN VALUES AND MAKE THEM BENEFICIAL TO SOCIETY

- Government will support the research of the beneficial use of AI
- AI should be developed to align with human values and contribute to human flourishing
- Stakeholders throughout society should be involved in the development of AI and its governance

WE WILL PLAN FOR A FUTURE IN WHICH AI SYSTEMS BECOME INCREASINGLY INTELLIGENT

- Governance models should be developed for artificial general intelligence (AGI) and superintelligence
- AGI and superintelligence, if developed, should serve humanity as a whole
- Long-term risks of AI should be identified and planned for
- Recursively self-improving AI development should be disclosed and tightly monitored and controlled for risk

AI PRINCIPLES

INCLUSIVENESS

WE WILL GOVERN AI AS A GLOBAL EFFORT

- Global cooperation should be encouraged to ensure the safe governance of AI
- Government will support the establishment of internationally recognised standards and best practices in AI, and when they are established, shall adhere to them

WE WILL SHARE THE BENEFITS OF AI THROUGHOUT SOCIETY

- Development of AI systems will be matched by a response to its impact on employment
- AI will be used to help humans retain purpose and flourish mentally, emotionally and economically alongside AI
- Access to training, opportunity and tools should be made available
- Education should evolve and reflect the latest developments in AI, enabling people to adapt to societal change

WE WILL PROMOTE HUMAN VALUES, FREEDOM AND DIGNITY

- AI should improve society, and society should be consulted in a representative fashion to inform the development of AI
- Humanity should retain the power to govern itself and make the final decision, with AI in an assisting role
- AI systems should conform to international norms and standards with respect to human values and people rights and acceptable behaviour

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

AI PRINCIPLES

CHANGELOG

VERSION	STAGE	DATE COMPLETED	SUMMARY OF CHANGES
1.0	Internal	06/09/2018	First draft
1.1	Internal	06/09/2018	Design
1.2	Consultation	10/09/2018	Add elaboration under each principle
1.3	Consultation	10/09/2018	Changes following steering meeting 2: **Alignment to Humanity**
1.4	Consultation	25/09/2018	Changes following steering meeting 2 **Equality to Inclusiveness**
1.5	Feedback	09/10/2018	Changes following first round of feedback by public and private sector entities
1.6	Internal Review	30/12/2018	Minor changes to the wording of principles: Humanity, Inclusiveness

AI GUIDELINES

دبـــي الذكية
SMART DUBAI

DUBAI AI ETHICS GUIDELINES

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 organisations using this technology must consider, is urgently needed.

Dubai's Ethical AI Toolkit has been created to provide practical help across a city ecosystem. It supports industry, academia and individuals in understanding how AI systems can be used responsibly. It consists of principles and guidelines, and a self-assessment tool for developers to assess their platforms.

The Dubai AI Ethics Guidelines relate to Ethics principle in Dubai AI Principles:

"AI systems should be fair, transparent, accountable and understandable"

They offer tangible suggestions to help stakeholders adhere to the principle. They deliver detailed guidance and are arranged according to the four sub-principles of Ethics principle:

- We will make AI systems **fair**
- We will make AI systems **accountable**
- We will make AI systems **transparent**
- We will make AI systems as **explainable** as technically possible

The guidelines are non-binding, and are being drafted as a collaborative, multi-stakeholder effort, with full awareness of organisations' needs to innovate and protect their intellectual property. This is a collaborative process where all stakeholders are invited to be part of the dialogue. We would like to see the Dubai AI Ethics Guidelines evolve into a universal, practical and applicable framework informing ethical requirements for AI design and use. With these guidelines our aim is to offer unified guidance that is continuously improved in collaboration with our communities. The eventual goal is to reach widespread agreement and adoption of commonly-agreed policies to inform the ethical use of AI not just in Dubai but around the world.

SCOPE

This document gives guidelines for achieving the ethical design and deployment of AI systems in both the public and private sectors. Specifically it covers the crucial issues of Fairness, Accountability, Transparency and Explainability of the algorithms at the heart of AI systems. This document does not cover issues relating to employment, security or any other aspects of the governance of artificial intelligence besides those mentioned above.

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. See the Definitions section for a full definition of 'significant decision' and 'critical decision'.

DEFINITIONS

For the purposes of these guidelines, the following definitions apply:
AI developer organisation
An organisation which does any of the following:

- determine the purpose of an AI system;
- design an AI system;
- build an AI system, or:
- perform technical maintenance or tuning on an AI system

Note 1 to entry: the definition applies regardless of whether the organisation 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 organisation and the border force is an AI operator organisation.

AI OPERATOR ORGANISATION

An organisation which does any of the following:

- use AI systems in operations, backroom processes or decision-making;
- use an AI system to provide a service to an AI subject;
- is a business owner of an AI system;
- procure and treat data for use in an AI system, or:
- evaluate the use case for an AI system and decide whether to proceed

Note 1 to entry: this definition applies regardless of whether the AI system was developed in-house or procured.

Note 2 to entry: it is possible for organisations to be both an AI developer organisation and an AI operator organisation

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].

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

Note 1 to entry: 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

Note 2 to entry: a particular feature of AI systems is that they learn behaviour and rules not explicitly programmed in

[1] Consistent with ISO/IEC 2382:2015

EXAMPLE:

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:

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.

BIAS

(of a system)

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

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, be especially sensitive, have the potential to cause high loss or damage, is societally significant, or sets important precedent.
Note 1 to entry: 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.

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

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.

ETHICS

(as applied to AI)

The concepts of fairness, accountability, transparency and explainability.

Note 1 to entry: for the purposes of this document, ethics does not include privacy concerns, model accuracy (except insofar as fairness and redress are concerned, for example), employment, or any other AI-related issues besides those listed in the definition.

FUNCTIONAL UNIT

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

INDIVIDUALLY SIGNIFICANT DECISION

A decision which has the potential for significant impact on at least one individual's circumstances, behaviour 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.

NON-OPERATIONAL BIAS

(of a system)

Bias that is either:

1. not a design feature; or
2. not important in achieving the stated purpose of the system

[3] From ISO/IEC 2382:2015

SET OF SIGNIFICANT-AT-SCALE DECISIONS

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

Note 1 to entry: the decisions need not be individually significant in order to qualify, in aggregate, as a set of significant-at-scale decisions

Note 2 to entry: examples of areas which have a large impact on society 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.

SIGNIFICANT DECISION

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

SUBJECT OF AN ARTIFICIALLY INTELLIGENT SYSTEM

(also 'AI subject')

A natural person who is any of the following:

- an end-user of an AI system
- directly affected by the operation of or outcomes of an AI system, or:
- a recipient of a service or recommendation provided by an AI system

GUIDELINES

1.1. WE WILL MAKE AI SYSTEMS FAIR

1.1.1. Consideration should be given to whether the data ingested is representative of the affected population

1.1.1.1. AI developer organisations and AI operator organisations should undertake reasonable data exploration and/or testing to identify potentially prejudicial decision-making tendencies in AI systems arising from biases in the data

EXAMPLE:

Following a natural disaster, a government relief agency uses an AI system to detect communities in greatest need by analysing social media data from a range of websites. However those communities where smartphone penetration is lower have 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.

1.1.1.2. AI developer organisations and AI operator organisations 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

1.1.1.3. AI developer organisations 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

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

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

1.1.2.2. When evaluating the fairness of an AI system, AI developer organisations and AI operator organisations should consider whether AI subjects in the same circumstances receive equal treatment

EXAMPLE:

An organisation 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.

1.1.3. Significant decisions informed by the use of AI should be fair

1.1.3.1. AI developer organisations and AI operator organisations could consider formal procedures such as Discrimination Impact Assessments as a means of ensuring fairness

1.1.4. AI operator organisations should consider whether their AI systems are accessible and usable in a fair manner across user groups

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

1.1.5.1. Efforts could be made to include people from diverse demographic backgrounds in the development and deployment processes

1.1.5.2. AI developer organisations 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 development and deployment to confirm these assumptions

1.2. WE WILL MAKE AI SYSTEMS ACCOUNTABLE

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

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

1.2.1.2. AI operator organisations and AI developer organisations should consider designating individuals to be responsible for investigating and rectifying the cause of loss or damage arising from the deployment of AI systems

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

1.2.2.1. AI operator organisations should only use AI systems that are backed by respected and evidence-based academic research, and AI developer organisations should base their development on such research

1.2.2.2. AI operator organisations 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

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[4]. If the potential impact of incorrect decisions had been assessed, mitigation measures such a user-friendly review procedure could have been set up.

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

1.2.2.4. In designing AI systems to inform significant decisions, AI developer organisations 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[5]. If the data is updated after the first case is encountered then it would avoid causing inconvenience on subsequent visits.

1.2.2.5. AI developer organisations and AI operator organisations should consider tuning AI models periodically to cater for changes to data and/or models over time

1.2.2.6. AI operator organisations should consider whether AI systems trained in a comparatively static environment will display model instability when deployed in dynamic environments

[4] Cabinet Office (UK), Data Science Ethical Framework, Version 1.0, licensed under the Open Government Licence v3.0, p. 13
[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-qBZUXEUG4dVk909eSg5qvfbpNlRhzlefWPtBwbxY/edit

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[6].

1.2.2.7. AI operator organisations should consider working with their vendors (AI developer organisations) to continually monitor performance

1.2.2.8. AI operator organisations 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

1.2.3. SUSPENDED - AI systems informing critical decisions should be subject to appropriate external audit

1.2.3.1. 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

1.2.3.2. 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

SUSPENSION NOTICE:

Guideline 1.2.3. has been suspended until further notice. Reason: no external auditing mechanism has yet been established.

[6] Stoica, I. et. al.,2017, A Berkeley View of Systems Challenges for AI, p. 2, https://arxiv.org/pdf/1712.05855.pdf

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

1.2.4.1. AI operator organisations which use AI systems to inform significant decisions should provide procedures by which affected AI subjects can challenge a specific decision concerning them

1.2.4.2. AI operator organisations should consider such procedures even for non-significant decisions

1.2.4.3. AI operator organisations should make affected AI subjects aware of these procedures, and should design them in a convenient and user-friendly way

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

1.2.4.5. AI operator organisations should consider instituting an opt-out mechanism for significant automated decisions

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[7]. 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.

1.2.4.6. AI operator organisations could consider compensating affected AI subjects in the case of loss or inconvenience caused by incorrect automated decisions

1.2.4.7. AI operator organisations 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

[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

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

1.2.5.1. 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

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

1.2.6.1. AI developer organisations 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

1.2.6.2. 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].

1.2.7. AI operator organisations should understand the AI systems they use sufficiently to assess their suitability for the use case and to ensure accountability and transparency

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

1.2.7.2. AI developer organisations should consider notifying customers and AI operator organisations 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-8529-11e8-96dd-fa565ec55929

1.3. WE WILL MAKE AI SYSTEMS TRANSPARENT

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

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

1.3.1.2. To facilitate the above, AI developer organisations and AI operator organisations 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

1.3.1.3. Where possible given the model design, AI developer organisations 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].

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

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

[9] See IBM WatsonPaths

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.

1.3.2.2. 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.

1.4. WE WILL MAKE AI SYSTEMS AS EXPLAINABLE AS TECHNICALLY POSSIBLE

1.4.1. AI operator organisations could consider providing affected AI subjects with a high level explanation of how their AI system works

1.4.1.1. AI operator organisations 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

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].

1.4.1.2. For non-sensitive public sector use cases designed for the common good, AI operator organisations 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')

1.4.2. AI operator organisations 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

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

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

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].

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

[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

CHANGELOG

VERSION	STAGE	DATE COMPLETED	SUMMARY OF CHANGES
1.0	Consultation	05/09/2018	Initial draft
1.1	Consultation	06/09/2018	Amendments prior to circulation
1.2	Consultation	10/09/2018	Added definitions; reworded guidelines accordingly
1.3	Consultation	10/09/2018	Added examples
1.4	Consultation	30/09/2018	Reformatting of examples
1.5	Consultation	01/10/2018	Added definition of AI
1.6	Consultation	03/10/2018	Reformatted & swapped some examples
1.7	Feedback	09/10/2018	Incorporated first round of feedback
1.8	Feedback	09/10/2018	Edited examples; miscellaneous other edits
1.9	Feedback	10/10/2018	Miscellaneous edits
1.10	Internal Review	10/10/2018	Added bibliography
1.11	Internal Review	30/12/2018	Minor tweaks to wording, addition to introduction section

This document was produced by Smart Dubai in collaboration with Falcon and Associates.

BIBLIOGRAPHY

1. 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

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

3. 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

4. European Parliament. (2017, February 16). European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).Retrieved from: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//EN

5. 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

6. 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

7. 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

8. 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

9. The New York City Council. (2018). A Local Law in relation to automated decision systems used by agencies. Retrieved from: http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0

10. 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

11. 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-qBZUXEUG4dVk909eSg5qvfbpNlRhzlefWPtBwbxY/edit

12. 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/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

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

14. House of Lords Select Committee on Artificial Intelligence. (2018). AI in the UK: ready, willing and able?

15. House of Commons Science and Technology Select Committee. Algorithms in Decision Making.

دبـــي الذكيـة
SMART DUBAI