COVID briefs
BUILDING BACK BETTER: POST-PANDEMIC CITY GOVERNANCE

# AI ETHICS IN POLICY AND ACTION: CITY GOVERNANCE OF ALGORITHMIC DECISION SYSTEMS

**Andrea G. Rodríguez**
*Researcher & Project Manager, CIDOB*
*Lead Researcher, Global Observatory of Urban Artificial Intelligence (GOUAI)*

## TABLE OF CONTENTS

## Executive summary

Cities' use of artificial intelligence (AI) systems is set to become the new normal. AI can help cities manage transportation, provide better water and energy services, improve air quality or the speed of everyday bureaucracy. However, AI can have potentially negative effects on citizens' rights and freedoms. These can be identified early and mitigated by applying an ethical lens to the design and use of AI in cities.

The field of artificial intelligence ethics offers a set of guidelines on how to reduce the potential negative impact of AI systems on fundamental rights and freedoms and how to ensure alignment with human values throughout the system's lifecycle. However, there is no consensus about what role ethics should play while investing in algorithmic decision systems or what constitutes an ethical issue. Moreover, there is no international regulation that harmonises the ethical development of algorithmic tools.

This paper examines the most important documents addressing the topic of AI ethics produced by public and private stakeholders from different regions of the world. Cities can learn from each other and adapt these lessons to the necessities of the frontline applications affecting the lives of citizens. Indeed, approaches have been made in the past to the ethical application of disruptive technologies in cities. Through the Sharing Cities Declaration cities have agreed to mitigate the negative impact of the platform economy, while the Cities Coalition for Digital Rights has committed to a fair, ethical and rights-based digital transition.

While the European approach puts ethics at the centre of development, it is still risk-based and focuses on high-risk applications – those that may cause harm and damage fundamental rights. US companies treat ethics as a complement to the technical fixes to misalignment and misbehaviour. Other documents, such as those produced by the Chinese Beijing Academy or Artificial Intelligence, focus on the idea of "good AI" and sustainability, but rather than putting people first, they target the good intentions of the developer. Section 1 examines the approaches cities have made to AI ethics and introduces the approaches to artificial intelligence of Microsoft and IBM, two important global players in the field of artificial intelligence closely connected with smart city solutions.

Section 2 opens the analysis up to a wider multidisciplinary portfolio and finds that while there is no consensus on what should be considered "ethical", a number of topics are commonly repeated in documents from government institutions, the business community and think tanks and academia. The present paper proposes these topics be considered minimal ethical standards and expands on them, offering new insights. The proposed framework considers that the following topics should be explored in order to mitigate the potential negative impact of AI:

- Fairness and non-discrimination
- Transparency and openness
- Safety and cybersecurity
- Privacy protection
- Sustainability
- Accountability

A self-assessment guide for compliance with the principles of the framework is provided in the form of actionable questions that may also serve as a directory for capacity building. These questions are based on the self-assessment list for trustworthy AI systems released by the European Commission's High-Level Expert Group on Artificial Intelligence (HLEG AI 2019) and adapted to the city context and to the topics covered by each principle.

To finish, the paper concludes that the lack of consensus and regulation about artificial intelligence is an opportunity for cities to discuss and adopt basic guidelines and incorporate them into their strategies, creating an alternative model of governance for artificial intelligence systems. It nevertheless recommends periodical audits of the systems in order to enforce these principles and encourages cities to establish strong channels of communication with residents to identify possible variations in aligned behaviour.

## Introduction

Algorithmic decision systems (ADS) can become important instruments for efficiently managing the negative effects of the urban metabolism and improving residents' living standards. The United Nations projects that by 2050 68% of the global population will live in urban areas. The challenge to provide services to a growing urban population while mitigating the negative effects of ecological degradation are two fields of opportunity for the use of ADS and, specifically, artificial intelligence (AI).

ADS is "automation by means of algorithms of multiple processes which underpin the decision-making process, including the collection and processing of data, as well as the execution of decisions with little or no human intervention" (Restrepo Amariles, 2020: 275). These processes are algorithmic tools that may or not use artificial intelligence, a set of techniques by which the algorithm *learns* from training data to perform tasks more accurately.

Cities are data mines. They receive constant input from the flows of materials and goods, people, climate and infrastructure as a result of the relationship between the residents and the city – the city both as a geographical and local administrative unit. The integration of information technology, objects and devices at the individual and public-local levels and the role of residents in smart cities (Townsend, 2014: 15) make the debates around the ethical design, use and implementation of AI at the city scale highly important.

Debates around smart city ethics address the challenges of the use of resident's data, including the problem of surveillance. In fact, the ethical and legal problems arising from the predictability of human behaviour and the potential misalignment and misuse of ADS are some of the key concerns city officials face when implementing new technologies at the city level. Smart city ethics also addresses the role of the private sector in the procurement and control of these new technologies.

Ellen P. Goodman's (2020) review of the issue reveals that the concentration of power and the potential to reuse citizens' data have led cities to come together to tackle three "ethical" issues: the privatisation of data, technology and services; the platformisation of cities, which creates new societal arrangements and can hamper equal opportunity for the sake of efficiency; and city domination by malfunction due to cybersecurity risks and/or market control of big companies. Nevertheless, the field of ethical AI systems in cities remains unexplored.

In recent years the business community and institutions have become key partners in the development of technological ethical solutions. Companies participate in large events discussing the matter and contribute to the field by issuing guidelines and participating in expert groups. Thus, the business community is closely involved with AI ethics at a time where there is no consensus on what constitutes an ethical standard or even how to drive the development of algorithmic decision systems,[1] with the exception of *soft law* and/or recommendations designed to obstruct the procurement of AI systems that do not meet the individual ethical standards of the buyer.[2]

The lack of international regulation and the discrepancies in what should be considered ethical for public administrations are the focus of the present paper. It starts with a definition of AI for the common good as a guiding principle from which the proposed framework stems. As such, the next section examines benchmark literature from the field of artificial intelligence ethics from the point of view of operational guidelines and therefore examines flagship documents from international institutions, governments, the private sector and cities. The documents are classified by geography and sector.

The geographical scope is limited to the European Union, the United States and China, primarily because they are the areas investing most in artificial intelligence. Logically, as they concentrate the greatest amounts of investment, partnerships and research, they are the most likely to set the global standards on artificial intelligence ethics. Another reason for this focus is that this paper aims to be a useful reference work for the Cities Coalition for Digital Rights, whose members are mostly based in these geographical areas.

Section 2 analyses the approach to artificial intelligence ethics of Microsoft and IBM, due to their importance in the global supply chain of key urban technologies and their market power. It describes the ethical priorities most repeated in the documentation analysed and expands and reviews them, incorporating an urban lens to create a framework of minimal ethical standards. The framework seeks to extend the list of issues that draw most consensus among government, industry, academia and city actors. It does not, therefore, reflect on what *should be ethical* but rather takes into consideration the work already implemented by stakeholders and systematically widens and adapts it to cities' needs. Moreover, as the six points of the framework take in both technical and socio-political discussions around the topic of AI ethics, there is enough flexibility to fit most cases. After the framework has been presented, the appendix AI Ethics in Action offers a set of key guiding questions for self-assessment and capacity building for cities to develop their own toolkits while implementing these principles.

1. A thorough examination of the debates around regulation can be found in: Feijóo et al. "Harnessing artificial intelligence (AI) to increase wellbeing for all: the case for a new technology diplomacy". *Telecommunications Policy, 44*, 2020.
2. This is the case for the City of Barcelona, whose artificial intelligence strategy provides clear guidelines for the administration to consider when jointly developing or purchasing an AI system to be used in the city.

Finally, the paper concludes that the lack of consensus and regulation offers an opportunity for cities and that the proposed framework complements the existing vision of AI cities developed in the Sharing Cities and the Cities Coalition for Digital Rights declarations.

## 1. Drafting artificial intelligence ethics in regulations and guidelines

The field of artificial intelligence ethics is characterised by a lack of consensus among stakeholders about what role ethics should play in the system lifecycle and what should be considered an ethical principle. There is interesting divergence on both these questions when it comes to the approach to AI development in different geographies and sectors. To present these differences, this section analyses working documents in the European Union, the United States and China. Furthermore, this part of the paper also considers the role of the private sector by examining the ethical guidelines issued by IBM and Microsoft, two key players in the artificial intelligence market that have a close relationship with cities and are often providers of smart city solutions (Townsend, 2014).

In the context of the European Union, the institutions have worked on a rich portfolio investigating the possible ethical pitfalls and legal loopholes by which the fundamental freedoms and rights of citizens could be compromised. Three documents are especially notable when it comes to artificial intelligence ethics in the EU: the 2019 report *Ethics guidelines for trustworthy AI*, written by the European Commission's High-Level Expert Group on Artificial Intelligence, the 2020 *White Paper on Artificial Intelligence*, and the 2021 proposal for an artificial intelligence act.

*Ethics guidelines for trustworthy AI* (EU HLEG AI, 2019) creates a framework of analysis of artificial intelligence systems. The objective of the proposed framework is to assess the level of trustworthiness of an AI model based on three principles: lawfulness, ethics, and technical and social robustness. When it comes to ethics, it proposes to pay special attention to vulnerable groups and offers a set of seven ethical guidelines to mitigate the risks posed by AI systems during the development and deployment phases. These requirements for trustworthiness were later expanded into a self-assessment list for institutions and developers working with artificial intelligence. It posed questions covering the topics of human agency, robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well-being, and system accountability (EU HLEG AI, 2019b).

The report was key for the development of a *White Paper on Artificial Intelligence* that integrated the group's recommendations and became the basis for the proposal for an artificial intelligence act (European Commission, 2021) released in April 2021 and under discussion in the European institutions as of November 2021. The white paper categorised the ethical principles discussed in the group's report into key requirements to achieve an ecosystem of excellence and trust, the two pillars of artificial intelligence development in the EU. It also

briefly touched upon uses of artificial intelligence that may constitute high-risk applications due to their adoption in sensitive sectors and their impact on citizens' rights and freedoms (European Commission, 2020). The AI act proposal confirms the European Union's risk-based approach to artificial intelligence.

On the other side of the Atlantic the approach to AI ethics enters market dynamics from the point of view of ensuring fairness in the development phase and training officials in its ethical use. Most of the documents that deal with AI ethics are developed by bodies of national security, such as the military or the intelligence community. *Preparing for the future of artificial intelligence*, a report published by the National Science and Technology Council under the Obama administration, is one such example.

The report complements national efforts to invest in strategic uses of artificial intelligence. Under this approach, the document analyses benign applications of artificial intelligence, especially in the fields of transportation and criminal justice, and provides a set of recommendations to mitigate possible negative effects. The document considers the role of ethics as a necessary element for the development of AI but, while this was the driving vector of the EU regulatory proposals, the United States emphasises technical over socio-philosophical approaches to AI ethics. As a result, ethics appears as a complement to technical solutions to misbehaviour to "put good intentions into practice" (National Science and Technology Council, 2016: 3).

Similarly, the National AI Initiative, an umbrella programme coordinating US advances in guidelines and strategic inputs for the development of the AI industry and the adoption of AI tools in the public sector, makes "trustworthiness" one of its strategic pillars. The term is defined as "accuracy, explainability and interpretability, privacy, reliability, robustness, safety and security, or resilience to attacks—and ensure that bias is mitigated" (National AI Initiative Office, 2021). The Initiative offers a set of assessment tools and technical standards for AI developers to advance on trustworthiness and ensure interoperability. But, as with the 2016 document, it prioritises technical solutions to philosophical considerations, although it does encourage public debate and multistakeholder engagement.

Probably the most socially aware document under the Initiative is the General Service Administration's (GSA) *Guide to AI Ethics*, which includes a series of questions for developers aimed at raising awareness around ethical issues such as diversity, representativeness and explainability. These guidelines, however, fall short when it comes to the cybersecurity of the tools and their safe use (GSA, 2020).

Although the United States prioritises technical fixes over ethical solutions, some interesting overlaps remain with the European approach, particularly when it comes to the role of technical robustness and fairness, and transparency. However, the focus is placed on boosting the development of market-ready applications in order to maintain economic competitiveness in the context of globalisation.

The artificial intelligence market can be divided into AI applications, AI platforms, AI system infrastructure and AI application development and deployment (IDC, 2021). Under this classification two American firms top the charts: Microsoft, for applications, system infrastructure and application development; and IBM for software platforms, applications and system infrastructure.

Microsoft has released a set of artificial intelligence principles whose objective is to create "responsible artificial intelligence […] that put people first" (Microsoft Corporation, 2019). Under this lens, Microsoft commits to the development of fair AI systems that perform reliably and safely. In addition, the company embraces the principles of privacy and data security, inclusiveness, transparency and accountability. Nevertheless, Microsoft understands transparency as the way to guarantee explainable artificial intelligence, in other words, the capacity for external investigators to *see through* the system's process of artificial reasoning to understand inaccuracies and malfunction.

Understanding transparency as AI explainability is a common topic in the business community. Another example can be found in IBM's *Everyday ethics for AI*. The company defines ethical principles as "specific virtues that AI systems should possess [and] guidance for designers and developers in training and building AI" (IBM, 2021). IBM understands that their artificial intelligence systems must be aligned with human values, be accountable and explainable, stick to algorithmic fairness principles and respect user data rights. IBM's approach to artificial intelligence ethics respects essential principles that match the key guidelines of the National Science and Technology Council, while Microsoft approach to AI ethics is more comprehensive and mimics the European Union's recommendations for the development of artificial intelligence.

Another interesting example of artificial intelligence ethics can be found in the Beijing AI principles. Developed by the Beijing Academy of Artificial Intelligence (BAAI) the principles offer a set of good practices at the development and implementation stages. For the BAAI, artificial intelligence ethics includes algorithmic fairness, non-discrimination and biases, transparency, explainability and predictability, and accountability and traceability. The six BAAI categories fall somewhere in between the US and European Union's approaches. However, the most interesting part of this document is the set of good practices that come with the section on ethics.

The BAAI recommends that AI should "be designed and developed to promote the progress of society and human civilization, to promote the sustainable development of nature and society, to benefit all mankind and the environment, and to enhance the well-being of society and ecology" (Beijing Academy of Artificial Intelligence, 2019). The Beijing AI Principles are based on the notion of responsibility and sustainability and, while ethics is understood as part of the design process, at the use-level the BAAI introduces a novel approach to artificial intelligence not seen in the previously analysed documents. The principles call for informed consent of stakeholders when it comes to the impact of the systems on their rights and interests in order to enhance transparency, but they do not contemplate the right of the citizen to opt out of the AI system lifecycle.

## 2. AI ethics in action: Towards a policy-oriented framework for AI ethics in cities

The principles examined below constitute a framework of minimal ethical standards, which also considers other, initially disregarded categories. "Diversity" thus becomes fundamental to the principle of "fairness and non-discrimination" and "human oversight, control, and auditing" are key to defining the use of artificial intelligence for the common good (see: section 3).

The first section provided an overview of the different approaches to artificial intelligence ethics by a selection of relevant stakeholders. It drew attention to the lack of consensus on what should be considered an ethical principle and what role ethics should play in the development and use of AI systems. This section expands the number of documents examined and offers a classification of ethical standards by means of repetition, that is, it presents an analysis of the topics that show most agreement between stakeholders and classifies them from most important for stakeholders to least important to stakeholders. With the distilled priorities, a framework to provide ethical guidance on the development of AI systems in cities is proposed. To do this, six principles are identified and operationalised through guiding questions.

The *Oxford Dictionary of Philosophy* defines *ethics* as "the study of the concepts involved in practical reasoning: good, right, duty, obligation, virtue, freedom, rationality, choice. Also, the second-order study of the objectivity, subjectivity, relativism, or scepticism that may attend claims made in these terms" (Blackburn, 2016). The objectivity and rightfulness of algorithms must be considered for their design to be ethical, along with the protection they provide to fundamental rights and freedoms when in use.

Table 1 offers a thorough examination of the ethical issues discussed not only in the documents reviewed in section 1, but also of other papers that have had a relevant impact in the development of the field, whether in terms of influence or number of citations (Hagendorff, 2020). The table shows a total of 19 topics discussed in 20 key documents from the public sector and international institutions (light blue, 7), the private sector (light orange, 7), and the think tank community and academia (light yellow, 6).

It should be noted that not all the 19 ethical principles shown in the table fit the definition of *ethics* provided above, but they have been treated as fundamental to guaranteeing the good use of artificial intelligence and/or showing the dichotomy between *good* development and *good* use. For example, "science-policy link" refers to the common development of good practices by the science community and the policy community together, which promotes the strengthening of the relationship between the two sets of actors as a means to guarantee the ethical development of artificial intelligence solutions.

Despite the lack of consensus over the role of ethics, some topics are constantly repeated in the different documents. These oft-repeated principles arguably constitute an ethical framework on whose importance and relation to AI ethics partial consensus exists. In order, these standards are *fairness and non-discrimination*, *safety and cybersecurity*, *privacy protection*, *accountability*, *sustainability*, and *transparency and openness*.

That they are the most repeated shows a level of general agreement about their importance, but none is a product of total consensus. 18 out of 20 documents examined agree that ethical AI systems should consider *fairness and non-discrimination*, the most commonly mentioned category, while *transparency and openness*, the least common category, constitutes an ethical principle for 15 out of 20 actors but 11 other actors mention

*explainability and interpretability*—important elements of algorithmic transparency—in their papers. The rest of the principles lie in between, with 16 papers mentioning *safety and cybersecurity*, *privacy protection and accountability,* and 15 citing *sustainability.*

The principles examined below constitute a framework of minimal ethical standards, which also considers other, initially disregarded categories. "Diversity" thus becomes fundamental to the principle of "fairness and non-discrimination" and "human oversight, control, and auditing" are key to defining the use of artificial intelligence for the common good (see: section 3).

**Table 1: Principles discussed in the documents examined by the author.**

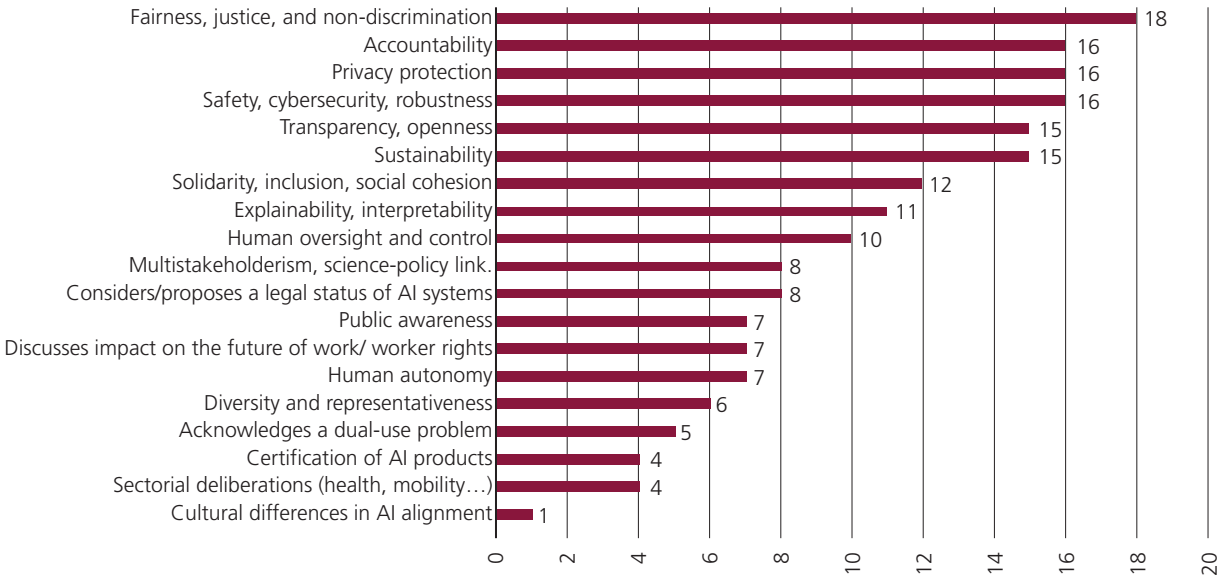
| | IEEE Ethically Aligned Design: A Vision… | EU Proposal for Laying Down Harmonised… | Beijing AI Principles | AI4People | Report on the Future of Artificial Intelligence | OECD Recommendation of the Council… | FLI Asilomar AI Principles | AI Now 2019 Report | Montréal Declaration of Responsible… | Sharing Cities Declaration | "Ethical Guidelines for Trustworthy AI" European… | Partnership on AI | Cities Coalition for Digital Rights Declaration | General Service Administration's Guide to AI Ethics | Microsoft AI Principles | IBM Everyday Ethics for Artificial Intelligence | DeepMind Ethics & Society Principles | Artificial Intelligence at Google | Principles for Accountable Algorithms… | OpenAI Charter | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Values & rights** | | | | | | | | | | | | | | | | | | | | | |
| Fairness, justice, and non-discrimination | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | | 18 |
| Accountability | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | | | ● | | 16 |
| Transparency, openness | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | | ● | | 15 |
| Sustainability | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | | | ● | ● | ● | | ● | 15 |
| Solidarity, inclusion, social cohesion | ● | | ● | ● | | ● | | ● | ● | ● | | ● | ● | ● | ● | | ● | | | | 12 |
| Explainability, interpretability | ● | ● | ● | ● | | | ● | | ● | | ● | | ● | ● | | ● | | | ● | | 11 |
| Diversity and representativeness | | ● | | | | | | ● | ● | | ● | | | ● | | | ● | | | | 6 |
| Cultural differences in AI alignment | ● | | | | | | | | | | | | | | | | | | | | 1 |
| **Security** | | | | | | | | | | | | | | | | | | | | | |
| Privacy protection | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | | | 16 |
| Safety, cybersecurity, robustness | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | | | ● | | ● | 16 |
| Public awareness | ● | ● | ● | ● | ● | | | | | | | ● | ● | | | | | | | | 7 |
| Acknowledges a dual-use problem | | ● | | | ● | | ● | ● | ● | | | | | | | | | | | | 5 |
| Certification of AI products | ● | ● | | ● | | | | | | | | | | | | ● | | | | | 4 |
| **Focus** | | | | | | | | | | | | | | | | | | | | | |
| Human oversight and control | ● | ● | ● | ● | | ● | ● | | | | ● | | | | | | | ● | ● | ● | 10 |
| Multistakeholderism, science-policy link. | ● | | ● | ● | ● | | ● | | | ● | | | | | | | ● | | | ● | 8 |
| Considers/proposes a legal status of AI systems | ● | ● | | ● | ● | ● | | ● | | ● | | | | | | | | | | | 8 |
| Discusses impact on the future of work/ worker rights | ● | | ● | | ● | ● | | ● | | ● | | ● | | | | | | | | | 7 |
| Human autonomy | ● | | ● | ● | | ● | ● | | ● | | ● | | | | | | | | | | 7 |
| Sectorial deliberations (health, mobility…) | ● | ● | | | ● | | | ● | | | | | | | | | | | | | 4 |
| Total ethical aspects | 17 | 14 | 13 | 13 | 11 | 11 | 11 | 11 | 11 | 10 | 10 | 8 | 8 | 7 | 6 | 6 | 5 | 5 | 5 | 4 | / |

Prepared by CIDOB.

The proposed framework is compatible with the work of cities to guarantee smart city ethics. The Sharing Cities Declaration agreed on ten principles, among which were data sovereignty, digital rights and city sovereignty. A brief mention of digital ethics appears in the seventh principle, where cities agreed to implement "ethical digital standards" that "include the rights of privacy, security, information self-determination and neutrality, giving citizens a choice of what happens to their digital identity, who uses their data online, and for which purposes" (Sharing Cities Action, 2018). Sharing Cities understood transparency as the means for citizens to understand their digital footprint, with the goal of empowering citizens to make decisions based on this knowledge, and added privacy and cybersecurity to the basket.

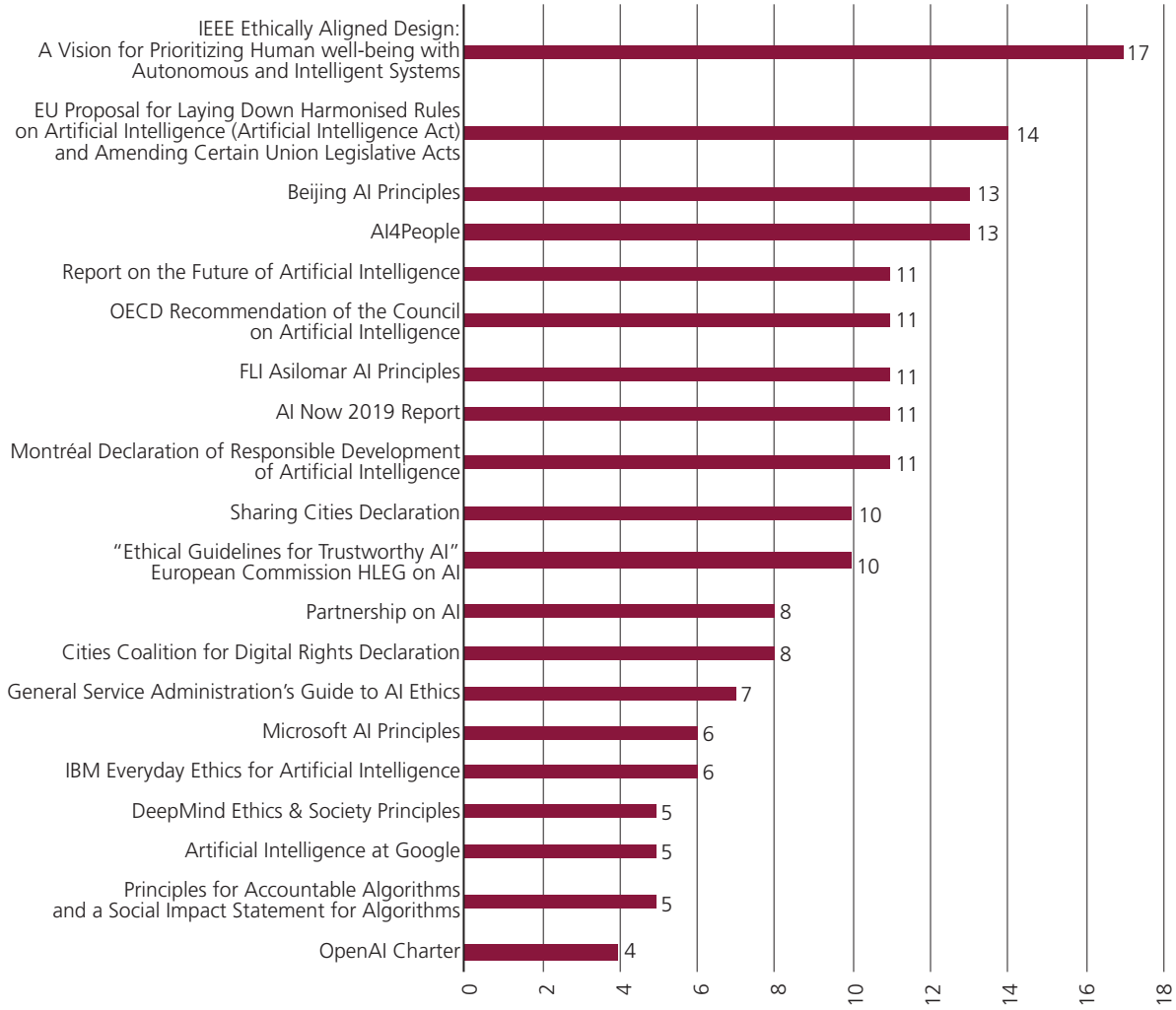
These three principles (transparency, cybersecurity and privacy) are at the heart of the Declaration of the Cities Coalition for Digital Rights, which also calls for citizen participation and universal access to the internet. However, the coalition reserves space in principles 3 and 5 for the realm of ethics and algorithms. Principle 3 calls for "understandable and accurate information about the technological, algorithmic, and artificial intelligence systems that impact their lives, and the ability to question and change unfair, biased, or discriminatory systems" (Cities Coalition for Digital Rights, 2018). Despite this, the mention of city ethics appears in principle 5, which encapsulates the idea of technological sovereignty mentioned in the Sharing Cities Declaration: "cities should define their own technological infrastructures, services and agenda, through open and ethical digital service standards and data to ensure they live up to the promise" (Sharing Cities Action, 2018). The principles of fairness and non-discrimination, the expansion of the notion of algorithmic accountability, and the introduction of a "green" lens in artificial intelligence systems would complement cities' understanding of smart city ethics.

**Figure 1: Number of ethical principles discussed in key documents from the public, private and academia/research sectors.**




Source: The author

**Figure 2: number of ethical principles discussed in each document investigated.**




Source: The author

## 2.1. Fairness and non-discrimination

The topics of fairness and non-discrimination in artificial intelligence systems often go hand in hand. The first deals with algorithmic justice, or how to mathematise what should be considered *fair* (Binns, 2018). The discussion of fairness is especially relevant while using AI applications in sensitive domains, such as in the allocation of resources (Cuellar & MacCoun, 2020) – although there is an ongoing debate around fairness. For that reason, it is easier to speak of "non-discrimination".

When developers' biases and prejudices are reflected directly in the model without these pre-existing conditions being challenged or mitigated the algorithmic tool may discriminate against certain groups or individuals. That is the reason the city of Nissewaard (Netherlands) stopped using AI to grant social assistance to children in need.

Nevertheless, having recognised discrimination as problematic, and given the legal implications of public administrations consciously discriminating against a group or individual (e.g. Niklas, 2020), the algorithmic non-discrimination literature tries to provide solutions for indirect discrimination ("disparate impact").

Disparate impact is a form of indirect discrimination that occurs when "neutral-sounding rules disproportionally affect a legally protected group" (Nunn, 2020: 186). Disparate impact can be a result of the training dataset or the architecture of the algorithm, but also of the changes in non-discriminatory practices by the algorithmic aggregation of new instances. In these circumstances, a previously certified "ethical" artificial intelligence system may create biased conclusions as it learns from the new input by a distributional shift or misbehaves by drawing conclusions from a limited sample ("scalable oversight").[3]

Moreover, the configuration of the training dataset can result in disparate impact upon minority groups. For that reason, the design of the AI system should consider the best way to represent minorities and protected groups in the dataset and assess the impact of misrepresentation or overrepresentation, that is, when data from these groups appear in the training dataset proportionally augmented or reduced to create balance between different groups.

*Guiding questions:*

- Does the system have a clear task to perform? How has the task been chosen? Have citizens participated in it in any way?
- What technology was used? Does it employ publicly available tools?
- Is there a strategy or a set of procedures for avoiding bias in the AI system during the design? Does it include diversity and representativeness? Has the system been tested by targeting problematic use-cases?
- *Connects with 2.2: is there a mechanism for citizens to report discriminatory or biased practices? Is there a strategy for investigating these issues?*
- *Connects with 2.3: is there a mechanism for monitoring the system lifecycle if the AI reproduces previously unnoticed biases?*

### 2.2. Transparency and openness

Transparency is the most elusive standard in the sample. The multiplicity of definitions given of the term reflect both the technical nature of AI and the socio-political context in which algorithmic decision systems are used.

From a technical point of view, transparency is often defined as explicability and interpretability. Interpretability refers to the ability of an AI model to determine the relation between two variables. An interpretable model would be able to understand which variable causes what effect. Interpretable models can also be explainable models. Explainability refers to the ability of an external auditor to understand how the technical architecture determines the weight

**3.** For a complete list of how things can go wrong and lead to discrimination see: Amodei et al., 2016. The text also offers suggestions for reaching a technical solution to the problem at the level of design.

of the parameters (Johnson, 2020). The technical understanding of algorithmic transparency is necessary to create methods for algorithmic accountability. Without explainability and interpretability, external actors would not be able to determine responsibility.

Additionally, from a socio-political point of view (Kaminski, 2020), this section relates to the transparency and administrative openness of AI systems *in use*, creating a direct relation with political accountability. It builds from the creation of an environment of trust between the city and its residents by promoting co-design and co-creation measures (see: section 3) and involving residents in the whole of the system's lifecycle. In this way, citizens can challenge the correlations (output) of the algorithm decision systems and help developers correct discriminatory practices. More citizen participation can also positively impact the legitimacy of the use of AI systems in the city and improve relations between residents and the city administration.

*Guiding questions:*

- Is there a mechanism for citizens to report potential abuses? How is the quality of the input assessed?
- Is the system traceable? Are the decisions of the system transparent?
- Is the system explainable? Do citizens understand the system's decisions? How do they communicate whether they understand them or not?

What mechanism does the city use to provide information to residents about AI systems deployed in the public space? How does the city provide information about risks and necessity of use?

### 2.3. Safety and cybersecurity

The alignment of artificial intelligence systems with human values and the capacity of the algorithm to remain aligned with its intention after deployment constitutes the key safety issue. Ensuring alignment is the overarching goal of artificial intelligence ethics and is directly related to predictability or how an AI system behaves in a certain environment (Cuellar & MacCoun, 2020).

Robustness – the ability of a system to function without errors – also involves ensuring a high degree of cybersecurity of the artificial intelligence system. Algorithmic tools may be vulnerable to adversarial attacks that seek to compromise the system's accuracy by creating noise in the training dataset or input perturbation (Akhtar & Mian, 2018).

Bad system performance can pose a risk to people's security. The City of Florence, for example, is developing a smart system at tram stops that recognises improper behaviour (Iolov, 2021). The system will be able to differentiate between people and vehicles and aims to increase safety by allowing trams to react earlier. Imagine, however, an adversarial attack that reduces the system's level of accuracy when identifying objects. In this case, people's safety could be compromised not by misuse, but by a low level of cybersecurity.

*Guiding questions:*

- What are the potential risks to the AI systems? Are there measures in place to mitigate potential negative consequences?
- Is there a mechanism to monitor and evaluate the performance of the AI system? Is the AI system reliable and stable? How does performance affect its behaviour?
- How is the system updated? Has the administration considered how to keep the system up-to-date?
- Does the design consider potential vulnerabilities related to data poisoning, model evasion or inversion?

### 2.4. Privacy protection

Privacy protection requires several levels of analysis. First, administrations must guarantee that the application is neither intrusive nor damages the fundamental right to privacy. Additionally, institutions and companies should ensure that data that relates to people is anonymous and should endeavour to protect the integrity of data throughout the lifecycle. Moreover, local administrations and companies should ensure that data is always available. This connects directly with providing the right of explanation (Cabral, 2021). Under European data protection law (GDPR), citizens have the right to learn about the performance of artificial intelligence and how a specific system works.

Second, data sovereignty is an important component of guaranteeing privacy protection. The Sharing Cities Declaration (2018) understands data sovereignty as the promotion of policies "in order that the personal data is controlled by citizens themselves, and are protected from being misused, collected of shared without explicit consent". For that reason, local actors should prevent companies reusing citizens' data without their explicit consent.

*Guiding questions:*

- What data has been used to train the algorithm? Where is that data from? What is the rationale behind the chosen data?
- Which measures have been put in place to achieve privacy-by-design systems?
- Who takes care of data? How is data processed? *What are the safeguards? (connects with 2.3.)*
- Does the system comply with data protection laws? How?
- What is the impact of the system on the individual's right to privacy, intimacy, dignity and integrity?

### 2.5. Sustainability

The use of artificial intelligence systems can negatively impact global efforts to fight climate change. Most of the documents examined agree that institutions and companies must find new ways to develop artificial intelligence systems in the most environmentally friendly way possible.

AI relies on storage in data centres that consume a lot of energy and are, most of the time, powered by non-renewable energy. Artificial intelligence applications may also increase the generation of electronic waste, as their use renders devices obsolete or requires the use of more devices. Moreover, the optimism over artificial intelligence can trigger a new form of competition between different administrations and companies to mine data that can be used to feed new systems or improve existing ones (Dauvergne, 2020).

Despite these negative effects, artificial intelligence systems can be very useful to local administrations working towards the UN's Sustainable Development Goals (EU HLEG AI, 2019), helping them make better-informed decisions by providing new insights that can contribute to making certain policies and actions more effective, for example by assisting cities in creating air pollution models. AI correlations can also help local administrations allocate resources better, promote the development of deprived areas with improved knowledge of their situations, and better understand cities' complexity.

*Guiding questions:*

- What are the potential environmental impacts of the design and deployment of the AI system? Does the assessment consider the whole lifecycle of the system or only its design?
- How can cities reduce, compensate for and mitigate these impacts? What is the role of the city during the data lifecycle? Do cities have urban data centres?
- What is the impact of the system on the local community (local economy, social interaction, resident-administration interaction, people's rights and freedoms, etc.)? *(Connects with principle 2.4.)*.

### 2.6. Accountability

The principle of accountability undoubtedly connects with the principles of fairness and transparency. Algorithms that are auditable, explainable and interpretable enable the exploration of the processes that may cause misbehaviour. Moreover, ensuring accountability in AI systems helps local governments comply with the right to explanation and helps create trust and good governance practices.

*Guiding questions:*

- *See principle 2*
- Is the system auditable? How frequently is the system audited *(connects with principle 3)*? Who audits the system? Will third parties do so or a taskforce from the local administration?

## 3. AI for the common good

A discussion of the benign societal implications of the implementation of AI should be a precondition for the design and use of artificial intelligence in cities. AI system design should reflect societal values and remain human-centric. In this sense, the 2017 Asilomar principles, the

result of the eponymous conference organised by the Future of Life Institute, state that AI "should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state of organization" (FLI, 2017). The conception of the common good thus justifies the issuance of minimal ethical standards that can be adopted by the maximum number of stakeholders and ratified by the citizenry.

One definition of human-centric AI is that the deployment of human-centric algorithmic decision systems should be conditioned by an evaluation of the benign impact of their use on people's lives, with consideration given to the malicious effects they may have on citizens' rights and freedoms. These malicious effects reflect the direct impact on rights (such as access to services or grants denied by biased AI systems) and indirect impacts, such as panopticon effects on behaviour due to an excess of monitoring or surveillance practices, or the reuse of data by unauthorised parties.

Another definition of human-centric AI recognises that by putting people at the centre of the design and use of artificial intelligence systems in cities the parties involved – in this case, residents – are those determining the social limits of AI systems ("human oversight and control"). For that reason, the principles of the framework described in section 2 consider citizen participation in the co-design of systems and co-creation of strategies of implementation as an alternative cycle of control. AI systems in the city should promote social inclusion and put extra emphasis on the fight against discrimination.

AI systems in the city should promote social inclusion and put extra emphasis on the fight against discrimination.

Citizens participate in the design of AI systems by becoming active players. With informed consent, citizens' data becomes an essential part of configuring training data sets, along with their participation in the experimentation phase. Moreover, in compliance with the principle of *fairness and non-discrimination*, citizens can report malperformance, discriminatory practices, drifts from the optimal accuracy rate of urban algorithms, or excessive exposure to these systems (principle 4). This, in turn, contributes to monitoring the systems' performance and ensuring AI safety (principle 3).

What is more, citizen participation during the system's lifecycle enhances the socio-political transparency that legitimises and justifies AI use in cities (principle 2) and helps create an environment of trust between city officials and residents, which in turn contributes to good governance and democratic participation – key concepts in the use of AI for the common good

What is more, citizen participation during the system's lifecycle enhances the socio-political transparency that legitimises and justifies AI use in cities (principle 2) and helps create an environment of trust between city officials and residents, which in turn contributes to good governance and democratic participation – key concepts in the use of AI for the common good.

## Conclusions and recommendations

The lack of consensus on what constitutes an ethical issue and the lack of international regulation on artificial intelligence (at least until the approval of the EU's Artificial Intelligence Act) provides an opportunity for cities. Under the umbrella of the Cities Coalition for Digital Rights, members have the chance to endorse a set of principles and put them into practice as means of alternative international algorithmic governance.

However important the role of ethics is for the correct development and use of AI systems, the risk management approach provides an interesting complement. AI ethics is a flexible list of principles that varies depending on the area of application and stakeholder involved. However, as the literature shows, the six principles analysed above are fundamental for the protection of citizens' rights and freedoms and ensuring algorithmic accountability and alignment.

These principles (*fairness and non-discrimination, safety and cybersecurity, accountability, privacy protection, sustainability*, and *transparency and openness*) are not independent from each other. In fact, each of them affects the rest. As an example, algorithmic *transparency* contributes to *accountability* because it allows external actors to identify the problem and allocate responsibility. *Fairness* and *non-discrimination* contribute to *privacy* as they guarantee the quality and integrity of the data used to feed the model. It could even be argued that *sustainability* forms part of *fairness*, as it contributes to social justice.

But as well as complementary, these principles can also be in tension with each other. For example, the protection of privacy can undermine the ability of companies and local governments to obtain the necessary data to ensure fairness. Cities will have to navigate these issues and, in cooperation with citizens, establish the means to ensure that trade-offs remain in line with citizens' interests, while preserving fundamental rights and freedoms. Using existing or new means of communication with citizens will be essential to channelling priorities while navigating trade-offs. In other words, only by ensuring constant interaction and by putting citizens' rights and interests first will cities be able to enforce human-centric – or humanistic – AI.

Lastly, to guarantee long-term alignment and correct behaviour, it is important to establish mechanisms for enforcing these principles after deployment. Enforcement should include oversight to ensure the readiness and fitness of the algorithmic tools after deployment. It is also advisable to establish channels of communication with citizens to ensure a quick response to bad performance or potential discriminatory practices and to create trust. As mentioned above, by using their advantage of proximity cities will be able to digitally advance and ensure that that the algorithms developed and used in the city remain aligned with the needs and interests of the citizenry. In other words, with people at the centre.

## References

Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning Computer Vision: A Survey. *IEEE Access, 6*. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8294186

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete Problems in AI Safety. *CoRR, abs/1606.06565*.

Barcelona City Council. (2021, April). *Government measure for a municipal algorithms and data strategy for an ethical promotion of*

*artificial intelligence*. Retrieved from https://ajuntament.barcelona.cat/digital/sites/default/files/mesura_de_govern_intel_ligencia_artificial_eng.pdf

Beijing Academy of Artificial Intelligence. (2019). *Beijing AI Principles*. Retrieved from https://www.baai.ac.cn/news/beijing-ai-principles-en.html

Binns, R. (2018). Fairness in Machine Learning: Lessons form Political Philosophy. *Proceedings of Machine Learning Research*, *81*.

Blackburn, S. (2016). *The Oxford Dictionary of Philosophy* (3rd ed.). Oxford: Oxford Univesity Press.

Cabral, T. S. (2021). Chapter 2: AI and the Right to Explanation: Three Legal Bases under GDPR. In D. Hallinan, R. Leenes, & P. De Hert (Eds.), *Data Protection and Privacy. Data Protection and Artificial Intelligence.* Oxford: Hart Publishing.

Cities Coalition for Digital Rights. (2018). *Declaration of Cities Coalition for Digital Rights*. Retrieved from https://citiesfordigitalrights.org/declaration

Cuellar, M. F., & MacCoun, R. J. (2020). Arguing over Algorithms: Mapping the Dilemmas Inherent in Operationalizing "Ethical" Artificial Intelligence. In W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms.* Cambridge: Cambridge University Press.

Dauvergne, P. (2020). *AI in the WIld: Sustainability in the Age of Artificial Intelligence.* Cambridge, MA: MIT Press.

European Commission. (2020). *White paper on Artificial Intelligence — A European Approach to excellence and trust.* Brussels.

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Brussels.

European Commission High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from Shaping Europe's digital future: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission High-Level Expert Group on AI. (2019). *Trustworthy AI Assessment List*. Retrieved from https://web.archive.org/web/20210504183552/https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Feijoo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R., . . . Xia, J. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: the case for a new technology diplomacy. *Telecommunications Policy, 44*.

Future of Life Institute (FLI). (2017). *Asilomar AI Principles*. Retrieved July 20, 2021, from https://futureoflife.org/ai-principles/?cn-reloaded=1

Goodman, E. P. (2020). Smart City Ethics: How "Smart" Challenges Democratic Governance. In M. D. Dubber, F. Pasquale, & S. Das (Eds.),

*The Oxford Handbook of Ethics of AI.* New York: Oxford University Press.

GSA. (2020). *CoE Guide to AI Ethics*. Retrieved from coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*(30).

IBM. (2021). *Everyday ethics for AI*. Obtenido de https://www.ibm.com/design/ai/ethics/everyday-ethics/

IDC. (2021, February 23). *IDC Forecasts Improved Growth for Global AI Market in 2021*. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS47482321

IMD Institute. (2020). *Smart City Index*. Obtenido de https://www.imd.org/smart-city-observatory/smart-city-index/

Iolov, T. V. (2021, June 03). *Florence traffic is about to get smarter*. Retrieved from The Mayor: https://www.themayor.eu/en/a/view/florence-traffic-is-about-to-get-smarter-8084

Johnson, J. (2020, July 16). *Interpretability vs. Explainability: The Black Box of Machine Learning*. Retrieved from BMC Blogs: https://www.bmc.com/blogs/machine-learning-interpretability-vs-explainability/

Kaminski, M. E. (2020). Understanding Transparency in Algorithmic Accountability. En W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms.* Cambridge: Cambridge University Press.

Microsoft Corporation. (2019). *Responsible AI*. Obtenido de https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

National AI Initiative Office. (2021). *Advancing Trustworthy AI*. Retrieved from National AI Initiative: ai.gov/strategic-pillars/advancing-trustworthy-ai/

National Science and Technology Council. (2016, October). *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

Niklas, J. (2020). Human Rights-Based Approach to AI and Algorithms. En W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms.* Cambridge: Cambridge University Press.

Nunn, R. (2020). Discrimination in the Age of Algorithms. In W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms.* Cambridge: Cambridge University Press.

Raymond, A. H., & Connelly, C. (2020). Governance of Algorithms: Rethinking Public Sector Use of Algorithms for Predictive Purposes. En W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms.* Cambridge: Cambridge University Press.

Restrepo Amariles, D. (2020). Algorithmic Decision Systems: Automation and Machine Learning in the Public Administration. In W. Barfield (Ed.), *The Cambridge Handbook of the Law of Algorithms.* Cambridge: Cambridge University Press.

Sharing Cities Action. (2018). *Sharing Cities Declaration*. Obtenido de https://www.sharingcitiesaction.net/wp-content/uploads/2019/05/Sharing-Cities-Declaration-1.pdf

Sustainable Cities Platform. (2016). *The Basque Declaration: New Pathways for European Cities and Towns*. Retrieved from https://sustainablecities.eu/fileadmin/repository/Basque_Declaration/Basque_Declaration_English.pdf

Townsend, A. M. (2014). *Smart Cities: Big Data, Civic Hackers, and the Quest for the New Utopia.* New York: W.W. Norton & Company.

U.S. Department of Homeland Security. (2021, August). *S&T Artificial Intelligence & Machine Learning Strategic Plan*. Retrieved September 13, 2021, from dhs.gov/sites/default/files/publications/21_0730_st_ai_ml_strategic_plan_2021.pdf

## APPENDIX: AI ethics in action: Operationalising the framework

### Fairness and non-discrimination

Does the system have a clear task to perform? How has the task been chosen? Have citizens participated in it in any way?

What technology was used? Does it employ publicly available tools?

Is there a strategy or a set of procedures for avoiding bias in the AI system during the design? Does it include diversity and representativeness? Has the system been tested by targeting problematic use-cases?

*Connects with principle 2: is there a mechanism for citizens to report discriminatory or biased practices? Is there a strategy for investigating these issues?*

*Connects with principle 3: is there a mechanism for monitoring the system lifecycle if the AI reproduces previously unnoticed biases?*

### Transparency and openness

Is there a mechanism for citizens to report potential abuses? How is the quality of the input assessed?

Is the system traceable? Are the decisions of the system transparent?

Is the system explainable? Do citizens understand the system's decisions? How do they communicate whether they understand them or not?

What mechanism does the city use to provide information to residents about AI systems deployed in the public space? How does the city provide information about risks and necessity of use?

**Safety and cybersecurity**

What are the potential risks to the AI systems? Are there measures in place to mitigate potential negative consequences?

Is there a mechanism to monitor and evaluate the performance of the AI system? Is the AI system reliable and stable? How does performance affect its behaviour?

How is the system updated? Has the administration considered how to maintain the system up-to-date?

Does the design consider potential vulnerabilities related to 1) data poisoning, model evasion or inversion?

**Privacy protection**

What data has been used to train the algorithm? Where is that data from? What is the rationale behind the chosen data?

Which measures have been put in place to achieve privacy-by-design systems?

Who takes care of data? How is data processed? What are the safeguards? *(connects with 2.3.).*

Does the system comply with data protection laws? How?

What is the impact of the system on the individual's right to privacy, intimacy, dignity and integrity?

**Sustainability**

What are the potential environmental impacts of the design and deployment of the AI system? Does the assessment consider the whole lifecycle of the system or only its design?

How can cities reduce, compensate for and mitigate these impacts? What is the role of the city during the data lifecycle? Do cities have urban data centres?

What is the impact of the system on the local community? (local economy, social interaction, resident-administration interaction, people's rights and freedoms, etc.)? (Connects with principle 2.4.).

**Accountability**

*See principle 2*

Is the system auditable? How frequently is the system audited *(connects with principle 3)?* Who audits the system? Will third parties do so or a taskforce from the local administration?