Leandry Junior Jieutsa
Researcher on AI Governance in cities, UNESCO Chair in Urban Landscape

1. Introduction

Fairness and non-discrimination are core values of urban AI in people-centred smart cities. Increasing discussion among researchers and policymakers testifies to the growing importance of addressing bias and discrimination in AI systems. Fairness derives from moral judgment, i.e. the process by which individuals determine what is morally right or wrong (Weinkauf, 2023). Although AI offers many advantages for cities, its deployment puts the quest for a fair city to the test by creating or reinforcing discrimination and inequalities. Thus, integrating fairness and non-discrimination principles into the urban AI life cycle is crucial to ensure the well-being of individuals and communities in smart cities. Nevertheless, operationalising this principle remains complex and ambiguous. To achieve this, cities need to articulate their various roles in AI governance, whether they are developers of internal solutions, responsible for the deployment of external systems or regulators . This requires the adoption of a variety of mechanisms, including socio-technical innovation, the establishment of local standards for fairness in AI and procurement standards. In addition, urban legislation must be introduced to protect the most vulnerable and guarantee citizens the exercise of their digital rights. However, these measures require resources, which cities can mobilise by promoting cooperation and networking.

2. A fair AI system is bias-free and used responsibly

Fairness and non-discrimination are complex and critical concepts in contemporary society (Barocas et al., 2023a). According to the Cambridge Dictionary, "fairness" refers to the quality of treating individuals equally in a manner that is just or reasonable. It respects people both as individuals and as members of society. Three primary elements, articulated in distributive and socio-relational dimensions, constitute this concept that pertains to individuals or groups (Barocas et al., 2023b): fair equality of opportunity, right to justification, and equality in relationships (Giovanola and Tiribelli, 2022). A fair society necessitates considering each individual or group of individuals according to their specific characteristics and circumstances to ensure equitable treatment and outcomes (Giovanola and Tiribelli, 2022; Lyu et al., 2023). Thus, it incorporates the notion of non-discrimination, which implies that no one should be excluded. Vulnerable individuals or groups are most susceptible to discrimination.

Operationalising this principle remains complex and ambiguous. To achieve this, cities need to articulate their various roles in AI governance, whether they are developers of internal solutions, responsible for the deployment of external systems or regulators.

The emergence of disruptive technologies such as AI challenges the dimensions of fairness. Two main factors are involved in the context of discrimination in connection with AI, namely algorithmic biases and the utilisation of AI-based systems (Ferrara, 2023; O'Neil, 2016; Wachter et al., 2021).

The first factor, algorithmic bias, distorts the original training data or the AI algorithm, leading to skewed and potentially detrimental results (Holdsworth, 2023). These biases reduce the accuracy and potential of AI with varying degrees of impact depending on the application. There are two main categories of bias in AI: automation bias and bias by proxy (Barocas et al., 2023a; González-Sendino et al., 2023). Automation bias is the large-scale propagation through AI system processes of social and cultural biases deeply embedded in historical training data used to fuel the AI system. This category includes human bias, data bias, learning bias and deployment bias. Bias by proxy happens when unintentional proxies for protected variables (e.g. gender, race) allow biases to be inferred, despite efforts to exclude them from training data.

The second factor is the utilisation of AI-based systems. Indeed, when employed for profiling or social control, systems infringe upon digital rights (Calzada, 2021; Cugurullo et al., 2022). By collecting and utilising personal information, facial recognition technologies, for instance, violate the privacy and personal data of citizens (UN-Habitat, 2023). Digital rights are interpreted as existing human rights that must be protected in the context of digital technologies, as physical and digital spaces are increasingly intertwined (UN-Habitat, 2020).

Algorithmic fairness is predicated on interrelated variables (Weinkauf, 2023). An automated decision system is considered fair when it does not rely on sensitive data such as gender or religion, does not disadvantage minorities, and is utilised responsibly.

3. The AI dilemma: balancing between opportunities and impacts of AI systems in cities

Historically, urban planning has contributed to creating and reinforcing different forms of urban inequalities and discrimination (Fainstein, 2009; Hall, 2014). The most affected populations are notably minorities and the most vulnerable, which vary according to context. Consequently, numerous concepts have emerged, such as Henry Lefebvre's "right to the city" or the "just city" (Fainstein, 2009; Fincher and Iveson, 2012; Harvey and Potter, 2009; Lefebvre, 1968). These concepts aim to make cities more equitable, particularly through access to urban services and opportunities, for an improved quality of life.

The emergence of AI challenges the just city by bringing opportunities for more inclusive cities while also creating and reinforcing different forms of inequalities and discrimination. Indeed, AI systems possess the capability to filter and process substantial volumes of data connected to extensive networks and the urban environment. Consequently, they can enable complex decisions to be made autonomously or semi-autonomously (Marvin et al., 2022; Sherman, 2023; Yigitcanlar et al., 2021). Explainable AI (XAI) methodologies can assist municipalities in comprehending the calculation of equity and its improvement (Lyu et al., 2023). The implementation of AI facilitates enhanced citizen-municipality engagement and optimises service delivery, particularly for the most vulnerable populations.

Unlike traditional forms of discrimination, discrimination automated by algorithms is more abstract or opaque and unintuitive, subtle, intangible, difficult to detect and large-scale.

For instance, deep learning tools enhance spatial data management to optimise service delivery in disadvantaged neighbourhoods in Durban, South Africa. Generative AI facilitates participatory planning processes by generating urban scenarios in real time, thus enabling more inclusive urban planning that incorporates diverse perspectives. Furthermore, municipal chatbots, such as those implemented in Helsinki, Finland, or Saint-Lin-Laurentides, Canada, automate citizen interaction. This improves the management of public services, particularly for individuals unfamiliar with often complex administrative procedures, or those who face difficulties in accessing services in person.

However, as previously stated, AI systems and the emphasis on the economic competitiveness of cities challenge the just city by producing unfair and discriminatory outcomes. Moreover, unlike traditional forms of discrimination, discrimination automated by algorithms is more abstract or opaque and unintuitive, subtle, intangible, difficult to detect and large-scale (Kleinberg et al., 2018; O'Neil, 2016; Sanchez et al., 2024; Wachter et al., 2021).

For example, tax algorithms targeting "foreign-sounding names" and "dual nationality" led to thousands of racialised families being falsely accused of fraud in the Netherlands. Globally, predictive policing systems, like Clearview AI, raise privacy concerns while reinforcing bias (Dauvergne, 2022; O'Neil, 2016). In 2021, *Forbes* reported algorithmic bias in mortgage applications, with 80% of Black applicants denied. Similarly, *The Markup* (2021) found applicants of colour were 40-80% more likely to face loan denials, underscoring the discriminatory impact of AI.

Furthermore, the concentration of wealth in large cities, due to urban AI, leads to urban gentrification (Sanchez et al., 2024). Access, particularly to housing, for low-income populations is becoming increasingly difficult, if not impossible. Urban AI deployment policies thus contribute to reinforcing asymmetries between territories and urban inequalities.

AI systems therefore have significant impacts on cities and societies. This ambivalence raises the need for effective governance. Additionally, due to its opacity and the scale of its impact, it becomes challenging for affected individuals to defend themselves or assert their rights. This calls into question the right to non-discrimination enjoyed by citizens because algorithmic decision-making systems disrupt traditional legal remedies and procedures for detecting, investigating, preventing and correcting discrimination (Wachter et al., 2021).

> Due to its opacity and the scale of its impact, it becomes challenging for affected individuals to assert their rights […] because algorithmic decision-making systems disrupt traditional legal remedies and procedures for detecting, investigating, preventing and correcting discrimination.

4. Policy recommendations

According to UNESCO recommendations, AI actors must adopt an inclusive approach aimed at rendering the benefits of AI technologies available and accessible to all, taking into account the specific needs of different groups (UNESCO, 2023). At the city level, the implementation of fairness and non-discrimination in urban AI systems necessitates the articulation of the diverse roles that municipalities assume as developers of in-house solutions (albeit relatively infrequently due to financial and technical constraints), deployers and regulators. Enhancing the equity of AI systems additionally entails consideration of their entire life cycle, addressing various aspects throughout the design, development and implementation processes. Furthermore, the provision of effective solutions to disparities in AI system outcomes commences with the identification of their underlying causes.

4.1. General recommendations:

- **Define a strategy:** Cities must implement AI strategies that are structured around the principles of fairness and non-discrimination. These strategic documents enable cities to establish a robust foundation and conduct a precise assessment of their AI-related objectives. This approach is essential for planning the integration of AI to maximise its benefits while mitigating potential risks. These strategies should be developed through a participatory process and accompanied by action plans that delineate concrete measures to ensure the equitable integration of AI that leave no one behind.

- **Establish risk levels according to applications:** Cities must identify high-risk AI applications within their jurisdictions, taking into account existing disparities and inequalities in the territory. The identification of these high-risk applications should be followed by the implementation of protective mechanisms. Applications related to essential social services should be classified as high-risk and prohibited from operating with full autonomy. For instance, the City of San Jose has implemented an AI registry structured around a rigorous assessment of AI systems. This process involves a risk analysis, followed by a more comprehensive impact assessment, depending on the level of risk, all documented via an "Impact Sheet" and an "AI Fact Sheet".

4.2. Specific recommendations for cities as developers of in-house solutions

- **Emphasise inclusive socio-technical innovation.** Incorporate diverse non-technical stakeholders throughout the AI life cycle. According to UN-Habitat, this AI life cycle comprises five phases: framing, design, implementation, deployment and maintenance. If decisions in these various stages are predominantly made by technical actors or homogeneous groups, there is a significant risk that their biases will be integrated into the AI system. This risk is particularly pronounced if the tool is subsequently applied or generalised to broader population segments. Local governments must place greater emphasis on interdisciplinarity and multidisciplinarity that integrates social groups into the life cycle of urban AI.

- **Implement fairness techniques,** including: preprocessing data (which involves identifying and addressing biases in the data prior to model training); model selection (which focuses on utilising model selection methods that prioritise fairness); and postprocessing decisions (which involves adjusting the output of AI models to mitigate bias and ensure fairness) (Ferrara, 2023).

- **Enhance diversity in database construction across three dimensions: teams, data and models.** Establishing diverse, interdisciplinary teams and implementing ongoing training in fairness and ethics are crucial for minimising biases. Regarding data, enhancing the collection of sensitive attributes (e.g. sex, race, ethnicity) and documenting data-related decisions promotes transparency and facilitates addressing real-world inequalities. For models, providing open access to the community for testing, ensuring transparent documentation, and utilising explainable AI (XAI) can aid in identifying and mitigating biases, thereby ensuring equitable outcomes (González-Sendino et al., 2023).

- **Integrate compensatory correlation in AI systems.** As indicated by Giovanola and Tiribelli (2022), ensuring fair equality of opportunity in AI systems cannot be limited to eliminating discriminatory biases in the training data. Urban AI systems should be designed to consider existing inequalities in their context and incorporate mechanisms to compensate for them. For instance, in a city where disparities exist among communities or social groups, urban AIs must account for these disparities and implement compensatory measures. This may manifest in the form of personalised content, for example.

- **Integrate mitigation techniques in the AI life cycle.** Neutralise discriminatory effects in the data during the pre-training phase through methods such as resampling (altering the size of the data set that affects the distribution without transforming the data), fair representation (achieved by eliminating information that can associate an individual with a protected group), and re-weighting (utilised to transform the data by modifying the weight in the data set). During the training phase, employ regularisation and adversarial training, which are the most common methods for this purpose. Other emerging approaches include decentralised learning, fair linear regression, DeepFair, multimodal models and fairlet clustering. During the post-training phase, implement equalised odds, calibrated equalised odds, and reject option classification.

4.3. Specific recommendations for cities as deployers and regulators

- **Establish local standards for fair AI**. Discrimination and inequalities can manifest differently depending on the context, affecting individuals or social groups in various ways and at different scales. Therefore, cities must implement fairness standards for urban AI that consider these local specificities. These standards should incorporate general principles while integrating local considerations. The goal is to ensure that urban AI does not reinforce existing discrimination or create new forms of bias that disproportionately affect the most vulnerable. These standards should be developed in consultation with local communities and cover the entire AI life cycle.

- **Establish procurement standards for fair AI**. Cities must ensure that entities providing them with services align with fair AI principles. This requires establishing procurement mechanisms that oblige service providers to comply with the city's fair AI standards. Providers must meet compliance requirements regarding their algorithms if they are to be used by the city. For instance, San Jose-led GOV AI in the USA has adopted and introduced the aforementioned AI FactSheet for Third-Party Systems. It is a harmonised template for vendors to provide detailed information about their AI products, covering aspects such as system purpose, training data, model details, performance metrics, bias management, robustness and human-computer interaction.

- **Implement urban laws that ensure the right to justification**. This right allows individuals affected by an AI system to understand the reasoning behind an algorithmic decision, enabling citizens to comprehend and control how they are treated by these systems. When this right is not adequately respected, individuals must have the ability to challenge and modify the underlying parameters of the decision. Therefore, cities must consider, throughout the process, whether to deploy or withdraw an AI system, particularly if an individual's request for explanation cannot be fulfilled. This measure allows individuals facing discrimination to assert their digital rights.

- **Establish advisory bodies to investigate, prevent and mitigate potential malicious uses of AI**. Local governments should set up multidisciplinary advisory bodies that include community organisations, academia, businesses and other stakeholders. These bodies will play an audit role to limit AI-related discrimination. They will assess the city's AI models based on fairness metrics. Their evaluation will (1) identify potential biases that could affect fairness, (2) select metrics to measure the fairness of AI systems, and (3) mitigate the impact caused by these biases. Additionally, they will act as advisory bodies to guide cities in their actions and policies regarding fair AI.

5. Limitations

Achieving fairness in AI is complex. Interventions aimed at achieving fairness in urban AI can create tensions with the very objectives of the algorithms themselves. This implies that cities must adopt a compromise-based approach to balance gains and benefits, while prioritising the well-being of individuals and communities. However, this principle can seem abstract, leaving room for divergent interpretations, which complicates the operationalisation of success and impact measures (Sadek et al., 2024). Therefore, cities need to implement a local approach to operationalising fairness and non-discrimination in urban AI. This holistic approach considers the socioeconomic and cultural configuration of the city throughout the entire AI life cycle.

From a technical perspective, fair urban AI requires diverse human resources and adapted infrastructure (Du et al., 2023; Marvin et al., 2022; Yigitcanlar et al., 2020, 2023). This, in turn, necessitates significant financial investments (Bettoni et al., 2021). Additional costs are also needed for continuous training and education of staff and communities (Sadek et al., 2024; Varanasi, 2023). Cities must also anticipate legal and compliance costs, including audits and system adjustments to meet regulatory standards. These investments can represent substantial expenses, especially for small and medium-sized cities.

To overcome these limitations, cities can rely on networking. These networks provide opportunities for knowledge sharing, policy innovation and coordinated responses to global issues. Some examples are:

Cities Coalition for Digital Human Rights: a platform to promote an inclusive and democratic development of new technologies in cities.

City AI Connect: A global learning community and digital platform for cities to trial and advance the use of generative artificial intelligence to improve public services.

GovAI: A coalition composed of over 1,000 members and over 350 local, state and federal entities united in the mission to promote responsible and purposeful AI in the public sector.

AI4Cities: A project that enabled Helsinki, Amsterdam, Copenhagen, Greater Paris, Stavanger and Tallinn to challenge the market to come up with AI-based solutions to reduce CO2 emissions in their energy and mobility domains.

References

Barocas, S., Hardt, M., and Narayanan, A. "Classification". In: Fairness and Machine Learning: Limitations and Opportunities. Cambridge: MIT Press, 2023a.

Barocas, S., Hardt, M., and Narayanan, A. "Relative notions of fairness". In: Fairness and Machine Learning: Limitations and Opportunities. Cambridge: MIT Press, 2023b.

Bettoni, A. *et al*. "An AI adoption model for SMEs: A conceptual framework". *IFAC-PapersOnLine*, 2021, 54(1), p. 702–708.

Calzada, I. "The right to have digital rights in smart cities". Sustainability (Switzerland), 2021, 13(20).

Cugurullo, F., *et al*. "Urban AI in China: Social control or hyper-capitalist development in the post-smart city?". *Frontiers in Sustainable Cities*, 2022.

Dauvergne, P. "Facial recognition technology for policing and surveillance in the Global South: A call for bans". *Third World Quarterly*, 2022, 43(9), p. 2325–2335. Routledge.

Du, J. *et al*. "Artificial intelligence-enabled participatory planning: A review". *International Journal of Urban Sciences,* 2023, 28(2), p. 183-210.

Fainstein, S. "Planning and the Just City". In: Marcuse, P., ed. Searching for the Just City: Debates in Urban Theory and Practice. London: *Routledge*, 2009.

Ferrara, E. "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies". *Sci*, 2023, 6(1), p. 3.

Fincher, R., and Iveson, K. "Justice and injustice in the city". *Geographical Research*, 2012, 50(3), p. 231–241.

Giovanola, B., and Tiribelli, S. "Weapons of moral construction? On the value of fairness in algorithmic decision-making". *Ethics and Information Technology*, 2022, 24(1), p. 3.

González-Sendino, R. *et al*. "A review of bias and fairness in artificial intelligence". *International Journal of Interactive Multimedia and Artificial Intelligence*, 2023.

Hall, P. "Cities of Tomorrow: An Intellectual History of Urban Planning and Design Since 1880". 4th ed. *Wiley-Blackwell,* 2014.

Harvey, D., and Potter, C. "The Right to the Just City". In: Marcuse, P., ed. Searching for the Just City: Debates in Urban Theory and Practice. *Routledge*, 2009.

Holdsworth, J. "What is AI bias?". *IBM*. 2023.

Kleinberg, J. *et al*. "Discrimination in the Age of Algorithms". *Journal of Legal Analysis,* 2018, 10(2005), p. 113–174.

Lefebvre, H. The Right to the City. 1968.

Lyu, Y. *et al*. "IF-City: Intelligible fair city planning to measure, explain, and mitigate inequality". *IEEE Transactions on Visualization and Computer Graphics*, 2023.

Martinez, E., and Kirchner, L. "The secret bias hidden in mortgage-approval algorithms". *The Markup*, August 2021.

Marvin, S. *et al* . "Urban AI in China: Social control or hyper-capitalist development in the post-smart city?". *Frontiers in Sustainable Cities*, 2022, 4, p. 1030318.

O'Neil, C. "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy". First edition. *Crown*, 2016.

Rezende, I. N. "Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective". *New Journal of European Criminal Law,* 2020, 11(3), p. 375–389.

Sadek, M. *et al*. "Challenges of responsible AI in practice: Scoping review and recommended actions". *AI & SOCIETY*, 2024.

Sanchez, T. W., Brenman, M., and Ye, X. "The ethical concerns of artificial intelligence in urban planning". *Journal of the American Planning Association*, 2024, 0(0).

Sherman, S. "The Polyopticon: A diagram for urban artificial intelligences". *AI and Society*, 2023, 38(3), p. 1209–1222.

UNESCO. "Readiness Assessment Methodology: A Tool of the Recommendation on the Ethics of Artificial Intelligence". UNESCO, 2023.

UN-Habitat. "Mainstreaming Human Rights in the Digital Transformation of Cities: A Guide for Local Governments". United Nations Human Settlements Programme, 2020.

UN-Habitat. "Human Rights in the Digital Era. United Nations Human Settlements Programme", 2023, p. 1–56.

Varanasi, R. A. "'It is Currently Hodgepodge': Examining AI/ML Practitioners' Challenges During Co-production of Responsible AI Values". 2023.

Wachter, S., Mittelstadt, B., and Russell, C. "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI". *Computer Law & Security Review*, 2021, 41, p. 105567.

Weinkauf, D. "Privacy Tech-Know Blog: When Worlds Collide – The Possibilities and Limits of Algorithmic Fairness (Part 1)". *Office of the Privacy Commissioner of Canada,* April 5, 2023.

Yigitcanlar, T., Agdas, D., and Degirmenci, K. "Artificial Intelligence in Local Governments: Perceptions of City Managers on Prospects, Constraints, and Choices". *AI and Society,* 2023, 38(3), p. 1135–1150.

Yigitcanlar, T. *et al*. "Responsible Urban Innovation with Local Government Artificial Intelligence (AI): A Conceptual Framework and Research Agenda". *Journal of Open Innovation: Technology, Market, and Complexity*, 2021, 7(1), p. 1–16.

Yigitcanlar, T. *et al*. "Contributions and Risks of Artificial Intelligence (AI) in Building Smarter Cities: Insights from a Systematic Review of the Literature". *Energies,* 2020, 13(6), 1473.