npj Urban Sustainability

https://doi.org/10.1038/s42949-026-00394-1

Article in Press

The metrics trap: how technical sophistication masks social harm in urban AI systems

Received: 15 June 2025

Accepted: 15 April 2026

Cite this article as: Mashhadi Moghaddam, S.N., Cao, H. The metrics trap: how technical sophistication masks social harm in urban AI systems. *npj Urban Sustain* (2026). https://doi.org/10.1038/s42949-026-00394-1

Seyed Navid Mashhadi Moghaddam & Huhua Cao

The Metrics Trap: How Technical Sophistication Masks Social Harm in Urban AI Systems

Authors: Seyed Navid Mashhadi Moghaddam[1*], Huhua Cao[1]

[1]Department of Geography, Faculty of Arts, University of Ottawa, Ottawa, Canada.

[*]Corresponding Author: navid.mm@uottawa.ca

Abstract

This study examines the deployment of artificial intelligence in urban systems through a multiple case study analysis of 28 implementations selected from 157 AI deployments across six domains (2015-2024). We reveal how cities have fallen into a "metrics trap", pursuing technical accuracy that fails to achieve policy goals. Our analysis documents three critical patterns: (1) an "accuracy illusion" where impressive performance metrics mask fundamental failures, exemplified by ShotSpotter's 97% acoustic accuracy yielding only 9.1% crime-fighting effectiveness; (2) discriminatory feedback loops that transform historical bias into computational destiny, affecting millions through predictive policing and housing algorithms; and (3) successful community resistance movements from Toronto to Detroit proving technological determinism is a myth. Cases including Amsterdam's algorithm register, Seoul's participatory waste management, and NYC Health + Hospitals' bias mitigation reveal varied approaches to democratic AI governance. These findings suggest cities can pursue both technical capability and democratic accountability.

Introduction

Cities worldwide increasingly deploy machine learning algorithms to manage transportation networks, allocate housing, predict crime, monitor environmental conditions, optimize energy consumption, and deliver health services [1-3]. This proliferation has accelerated markedly: documented AI use cases in local governments nearly tripled from 22 cases in 2019 to 58 cases in 2020, peaking at 68 cases in 2021[4]. Between 2010 and 2020, AI tools became more accessible to municipalities as costs declined and commercial solutions expanded; the release of generative AI systems such as ChatGPT in 2022 further accelerated adoption, enabling diverse stakeholders to experiment with AI for urban analysis and service delivery[5]. Yet as algorithmic systems become embedded in urban infrastructure, a fundamental tension emerges between technical performance and policy effectiveness, a tension this study systematically examines through analysis of 157 documented implementations across 27 countries. Three interconnected scholarly domains frame our investigation: critical algorithm studies documenting systematic harms from automated decision systems [6-9]; urban governance research critiquing technocratic smart city approaches [10-13]; and science and technology studies analyzing how technological systems embed political values and reshape power relations [14-16]. While these literatures establish that algorithmic systems can perpetuate discrimination and reshape urban governance, they provide limited

empirical specification of the mechanisms through which technical metrics diverge from policy outcomes, how communities successfully contest algorithmic deployments, and what institutional arrangements enable democratic alternatives.

Understanding these dynamics requires distinguishing among forms of algorithmic opacity and transparency that are often conflated in policy discourse. Black-box models are algorithms whose internal decision-making processes remain opaque to human users, including deep neural networks and ensemble methods that learn patterns through parameters lacking meaningful human interpretation [17-19]. Post-hoc explainability methods such as LIME and SHAP attempt to approximate black-box reasoning after decisions are made, but as Rudin [20] argues, these provide "illusion of transparency without genuine interpretability", what Ananny & Crawford [21] term "seeing without knowing." Inherently interpretable models, by contrast, are designed from inception to be comprehensible, using transparent logic that humans can examine and contest [22,23]. Distinct from model interpretability is process transparency: the visibility of how algorithmic systems are developed, procured, deployed, and governed, including documentation of purposes, data sources, and accountability mechanisms [15,24]. This distinction proves crucial because a city may deploy interpretable models without transparent governance processes, or may publicly document opaque systems without enabling meaningful scrutiny of their logic. For this analysis, we classify implementations along a transparency spectrum based on three dimensions: model interpretability, which concerns whether the algorithm uses inherently comprehensible logic (e.g., decision trees, scoring systems, logistic regression) versus opaque architectures requiring post-hoc explanation; process transparency, which examines whether governance procedures including purpose, data sources, decision criteria, and accountability mechanisms are publicly documented; and contestability, which addresses whether affected stakeholders can understand, question, and influence algorithmic outputs.

Systems may exhibit transparency on some dimensions while remaining opaque on others; our Theme 4 cases were selected for demonstrating transparency across multiple dimensions rather than requiring perfection on any single criterion. Our analysis examines how these dimensions interact, revealing how their configuration shapes whether algorithmic systems serve or undermine democratic urban governance.

The widespread deployment of black-box systems in urban contexts rests on an assumption that merits empirical scrutiny: that achieving high predictive accuracy requires accepting algorithmic opacity. This supposed interpretability-performance tradeoff serves as justification for deploying inscrutable systems, with vendors claiming that transparency would sacrifice capability and officials accepting opacity as the cost of technical sophistication [25,26]. However, recent technical scholarship challenges this assumption. Rudin [20] argues the tradeoff is "not a real figure," providing evidence that interpretable models match black-box performance across diverse applications. Wagner et al.[27] demonstrated interpretable models explained 84% of variance in Berlin's urban emissions analysis while revealing policy-relevant threshold effects. Kim et al.[28] showed explainable approaches in Seoul's urban modeling maintained competitive accuracy while providing actionable insights. When the interpretability-performance tradeoff is accepted as given, it enables what we term the "metrics trap": algorithmic systems optimized for technical proxies that diverge from, or actively undermine, their stated policy purposes. This pattern extends Strathern [29]'s formulation of Goodhart's Law ("when a measure becomes a target, it ceases to be a good measure") into algorithmic governance, where Thomas and Uminsky [30] document how metric optimization can

become actively harmful. Muller [31] analyzes this as "the tyranny of metrics" across institutions; our study examines how it operates specifically in urban AI deployments.

Our analysis draws on theoretical frameworks from multiple disciplines to interpret patterns across cases. From science and technology studies, we employ Winner [16]'s insight that artifacts have politics, that technological systems embed values and enable particular governance arrangements, alongside Jasanoff [32]'s concept of co-production, whereby technical and social orders are constituted together. From algorithmic governance scholarship, we draw on Ensign et al.[33]'s formal models of feedback loops in predictive systems and Lum ans Isaac [34]'s analysis of how biased data generates biased predictions that produce biased enforcement. From institutional theory, we employ Ostrom [35]'s polycentric governance framework and (Fung & Wright [36]'s concept of empowered participatory governance. From social movement scholarship, we draw on resource mobilization theory [37], policy diffusion mechanisms [38,39], and frame alignment processes [40]. These frameworks enable us to move beyond documenting that algorithmic harms occur to explaining why harmful systems persist, how communities successfully resist them, and what institutional mechanisms enable democratic alternatives. We also distinguish transparency-as-disclosure from transparency-as-accountability [24], a distinction that proves essential for understanding why algorithm registers and documentation requirements vary in their governance effects.

Through multiple case study analysis of 28 in-depth cases selected from 157 documented implementations across six urban domains (2015-2024), this article examines how the current paradigm of urban AI development produces systematic divergence between technical metrics and policy outcomes. We document cases where high technical accuracy coexists with limited policy effectiveness, ShotSpotter achieving 97% acoustic accuracy while only 9.1% of alerts led to gun crime evidence [41]; COMPAS achieving aggregate accuracy while exhibiting racially disparate error rates [42]; healthcare algorithms accurately predicting costs while systematically under-identifying Black patients for care [43]. We examine how algorithmic systems can create feedback dynamics that amplify existing disparities [6,33,44]. We analyze conditions under which community resistance achieved policy changes, from Boston's facial recognition ban to Detroit's comprehensive reform [45,46]. And we document transparent alternatives, Amsterdam's algorithm register, Seoul's participatory waste management, NYC Health + Hospitals' bias mitigation, where interpretable systems and democratic governance processes produced documented benefits [47-49]. These patterns suggest that the choice between algorithmic capability and democratic accountability may be false: transparent systems in our sample achieved their documented outcomes not despite but through their responsiveness to democratic input. We offer these findings as systematic observations from a substantial sample warranting further investigation, while acknowledging that generalization requires validation across broader and more representative cases.

Results

Overview of the Case Universe

Our analysis draws from 157 documented AI implementations across six urban domains spanning 27 countries from 2015 to 2024 (Figure 1). This mapping enabled identification of patterns in urban AI adoption within our sample, though we acknowledge this dataset reflects documented and accessible cases rather than a comprehensive global census. The temporal boundaries capture the period following the deep learning revolution while providing sufficient time for implementation outcomes to become observable.

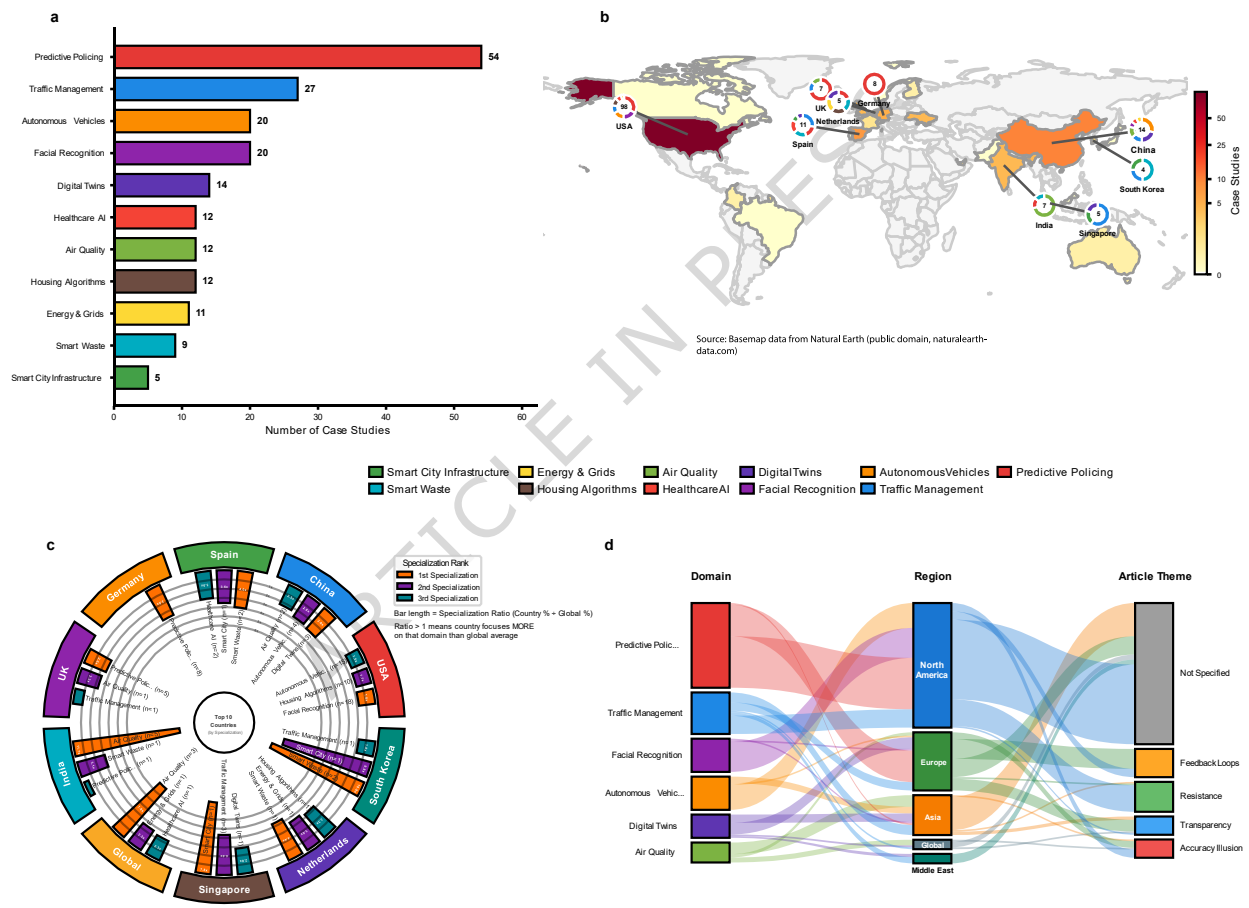

Figure 1 | Global landscape of urban AI implementations (2015–2024). **a**, Distribution of 157 documented cases across eleven urban domains, with predictive policing comprising the largest share (n=54). **b**, Geographic distribution showing concentration in North America and Europe, with the United States accounting for the most cases (n=98). **c**, Country-level domain specialization patterns for the ten most-represented nations, with bar length indicating specialization ratio relative to global distribution. **d**, Sankey diagram illustrating relationships between domains, regions, and analytical themes examined in this study. Case selection reflects documentation accessibility rather than comprehensive global deployment. For Panel B, basemap data from Natural Earth (public domain, naturalearthdata.com)

Within our sample, domain distribution (Figure 1a) shows notable variation. Predictive policing and public safety systems account for 54 cases (34.4%), followed by traffic management (27 cases, 17.2%), facial recognition (20 cases, 12.7%), and autonomous vehicles (20 cases, 12.7%). Domains associated with transparent governance approaches, including digital twins and algorithm registers (14 cases), smart waste management (9 cases), and smart city infrastructure (5 cases), constitute a smaller proportion. This distribution within our sample may reflect documentation availability and research attention rather than actual global deployment patterns; surveillance-related systems tend to generate more public scrutiny and thus more accessible documentation than routine municipal services.

Geographic distribution within our sample (Figure 1b) shows concentration in English-speaking and European contexts, with the United States accounting for the largest share, followed by European nations including Spain, Germany, and the UK. Asian implementations appear primarily from China, India, and Singapore. The Sankey diagram (Figure 1d) visualizes relationships between domains, regions, and article themes within our dataset. Country-level patterns (Figure 1c) show that within our sample, different nations appear more frequently in particular domains, for instance, cases from Germany cluster in predictive policing, while cases from India concentrate in air quality monitoring, and South Korean cases feature prominently in smart waste management. These patterns likely reflect both actual policy emphases and language/documentation accessibility limitations in our search strategy.

From this broader dataset, we selected 28 cases for in-depth thematic analysis based on documentation richness (evidence from at least four source types), demonstrated social impact, and theoretical significance for illuminating the metrics trap phenomenon. These cases were organized across four analytical themes: the accuracy illusion (Theme 1, examining seven cases), discriminatory feedback loops (Theme 2, seven cases), community resistance (Theme 3, seven cases including one contrast case), and transparent alternatives (Theme 4, seven cases). Some cases appear in multiple themes where their characteristics warranted examination from different analytical perspectives. The following sections examine these patterns in detail.

Theme 1: The Accuracy Illusion

The accuracy illusion emerges as a recurring pattern across urban AI implementations in our sample, where technical performance metrics diverge from policy outcomes. Our analysis of 21 cases across eight countries exhibiting this pattern (Figure 2) reveals five distinct manifestations: detection/prediction disconnected from real-world action (5 cases), systemic failures despite claimed accuracy (5 cases), vendor claims contradicted by verified reality (4 cases), accuracy metrics masking fairness failures (3 cases), and technical success without social impact (3 cases). Within our sample, predictive policing systems account for the largest share (9 cases), followed by healthcare AI (5 cases), with housing, autonomous vehicles, environmental monitoring, and traffic management comprising the remainder (Figure 2D). Notably, 30% of systems exhibiting accuracy illusion patterns in our sample have been ended or cancelled, while 65% remain active despite documented gaps between technical claims and policy effectiveness (Figure 2C). This section examines seven cases in depth to illustrate the mechanisms through which impressive technical metrics can obscure fundamental failures in achieving stated policy goals.

Figure 2 | The Accuracy Illusion: Evidence from 21 Case Studies Across 8 Countries. **a,** The metrics gap between technical claims and social outcomes, with color coding indicating comparable metrics (green), different metric types (orange), and systems that have ended (red highlight). ShotSpotter Chicago exhibits the largest documented gap (88 percentage points). **b,** Taxonomy of accuracy illusion manifestations showing five distinct patterns across cases. **c,** System outcomes: 30% of systems in our sample have been ended or cancelled, 65% remain active, and 5% exist only as research tools. **d,** Domain distribution within Theme 1 cases, with predictive policing representing the largest category (9 cases) followed by healthcare AI (5 cases).

The acoustic gunshot detection system ShotSpotter exemplifies the accuracy illusion through its divergence between technical precision and operational utility. The company claims 97% accuracy with 0.5% false positive rates in detecting gunshot sounds [50]. However, the Chicago Office of Inspector General's analysis of 50,176 ShotSpotter alerts from January 2020 through May 2021 found that only 9.1% led to evidence of a gun-related criminal offense [41], an 88 percentage point gap between claimed acoustic accuracy and crime-fighting effectiveness (Figure 2A). This disconnect stems from a category error: accurately detecting loud sounds differs from identifying gun crimes requiring police response. [51] analyzed ShotSpotter implementations across 68 large metropolitan counties from 1999-2016, finding no

significant impact on firearm homicide or arrest rates. A 2024 New York City audit revealed that 87% of ShotSpotter alerts were confirmed as false alarms, with fewer than 0.9% resulting in firearm recovery. As Piza[52] documented in Kansas City, while the technology achieves spatial precision in locating sounds, this technical achievement did not translate into measurable public safety improvements in that context. Chicago terminated its $33 million ShotSpotter contract in September 2024.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm illustrates how moderate overall accuracy can mask disparate error rates across demographic groups. ProPublica's analysis of over 7,000 criminal defendants in Broward County, Florida, found that COMPAS achieved approximately 60-70% overall accuracy in predicting recidivism [42]. However, this aggregate metric concealed differential false positive rates: Black defendants were falsely flagged as high-risk at 44.9% compared to 23.5% for white defendants. Conversely, white defendants who reoffended were incorrectly labeled low-risk 47.7% of the time compared to 28.0% for Black defendants. Even after controlling for criminal history, age, and gender, Black defendants remained 77% more likely to be scored as high-risk for violent recidivism. The accuracy illusion operates here through aggregation: overall performance metrics obscure how errors distribute across groups, a finding consistent with subsequent research demonstrating mathematical tensions between different fairness criteria in risk assessment [53,54]. Similarly, the Obermeyer healthcare algorithm achieved high accuracy for its optimization target, healthcare costs, yet only 17.7% of Black patients who should have been identified for extra care based on health needs were flagged, compared to an estimated 46.5% if the algorithm performed equitably [43]. The algorithm accurately predicted costs, but structural inequalities meant Black patients generated lower costs despite equal or greater health needs due to barriers in accessing care.

Environmental and health prediction systems in our sample exhibit a related pattern: technical accuracy that fails to capture differential impacts across populations. Delhi's air quality prediction system using Grey Wolf Optimization with Decision Tree algorithms achieved 88.98% accuracy for Air Quality Index prediction [55]. However, standardized monitoring networks may not capture hyperlocal variations affecting residents near industrial areas or informal settlements, where studies have documented pollution levels 30-40% higher than city averages. The system optimizes for meteorological prediction rather than equitable exposure assessment. Boston University researchers' hospitalization prediction model achieved 82% accuracy compared to 56% for clinical guidelines [56], yet the model's reliance on medical records and diagnostic data means it cannot incorporate social determinants, housing instability, food insecurity, transportation barriers, that shape health outcomes. For patients facing structural barriers, accurate prediction without addressing root causes documents outcomes rather than enabling intervention. COVID-19 contact tracing applications achieved 70-95% proximity detection accuracy in laboratory conditions, yet real-world effectiveness was limited by adoption rates below 20% in many countries and inability to distinguish transmission risk contexts [57]. Singapore's TraceTogether required 75% adoption for meaningful epidemic control, a threshold that was not reached.

These six cases suggest several mechanisms through which accuracy illusion may operate, though we note these patterns derive from our selected sample rather than representing universal claims about urban AI. First, each case exhibits misalignment between algorithmic optimization targets and policy objectives: ShotSpotter optimized for sound detection rather than crime prevention; COMPAS and Obermeyer optimized for aggregate prediction rather than equitable outcomes; environmental and health systems optimized for technical accuracy rather than population-level impact. Second, aggregated metrics can obscure distributional failures, overall performance statistics may hide how accuracy varies

across populations or contexts. Third, in several cases, technically accurate systems appeared to generate limited policy value: Chicago and New York audits found ShotSpotter alerts rarely led to actionable outcomes; COMPAS accuracy barely exceeded chance while exhibiting disparate error rates; high-performing health predictions could not address social determinants driving hospitalization. Table 1 summarizes these patterns. We emphasize that these observations derive from documented cases in our sample; whether they represent broader patterns in urban AI deployment requires further systematic investigation across larger and more representative samples.

Table 1. The Metrics Gap in Selected Urban AI Systems

Case	Domain	Technical Metric Claimed	Documented Social Outcome	Metrics Gap	Failure Mode
ShotSpotter (Chicago)	Public Safety	97% acoustic accuracy	9.1% of alerts led to gun crime evidence	88pp	Sound detection ≠ crime prevention
ShotSpotter (NYC)	Public Safety	High detection rate	87% false alarms; <0.9% firearm recovery	12pp documented	Lab accuracy ≠ field effectiveness
COMPAS (Broward County)	Public Safety	60-70% overall accuracy	2x false positive rate for Black defendants	0pp overall; disparate errors	Aggregate accuracy ≠ equitable outcomes
Obermeyer Healthcare	Health	High cost prediction accuracy	17.7% vs 46.5% Black patient identification	164% potential increase if corrected	Cost proxy ≠ health needs
Delhi Air Quality	Environment	88.98% AQI accuracy	30-40% higher exposure in informal settlements not captured	Unknown	City-wide accuracy ≠ hyperlocal equity
Boston Hospitalization	Health	82% accuracy (vs 56% guidelines)	Cannot address social determinants	26pp improvement without root cause intervention	Prediction ≠ prevention
COVID-19 Contact Tracing	Health	70-95% proximity accuracy	<20% adoption in many countries	~62pp gap between lab and field	Technical accuracy ≠

					population impact

Note: pp = percentage points. Metrics gaps calculated where comparable measures available. Singapore's TraceTogether is included as a documented example within the COVID-19 contact tracing case, illustrating how even well-resourced national implementations faced the adoption threshold barrier that characterized this technology globally.

Theme 2: Discriminatory Feedback Loops

Discriminatory feedback loops represent a pattern wherein algorithmic systems may not only reflect existing societal biases but potentially amplify them through iterative operation. Our analysis identified 11 cases across three domains exhibiting documented feedback loop characteristics: predictive policing (5 cases), housing algorithms (4 cases), and energy systems (2 cases) (Figure 3a). These cases span implementation periods from 2007 to present, with three systems subsequently terminated or discontinued (Chicago SSL in 2019, Palantir New Orleans in 2018, HART Durham in 2020), while others remain active or face ongoing regulatory action (Figure 3c). Regulatory responses to documented harms in these cases have resulted in over $36 million in fines and settlements, including $23 million against TransUnion, $4.2 million against AppFolio, and $3 million against RealPage (Figure 3d). This section examines seven cases in depth to illustrate mechanisms through which algorithmic systems may generate self-reinforcing cycles of disadvantage.

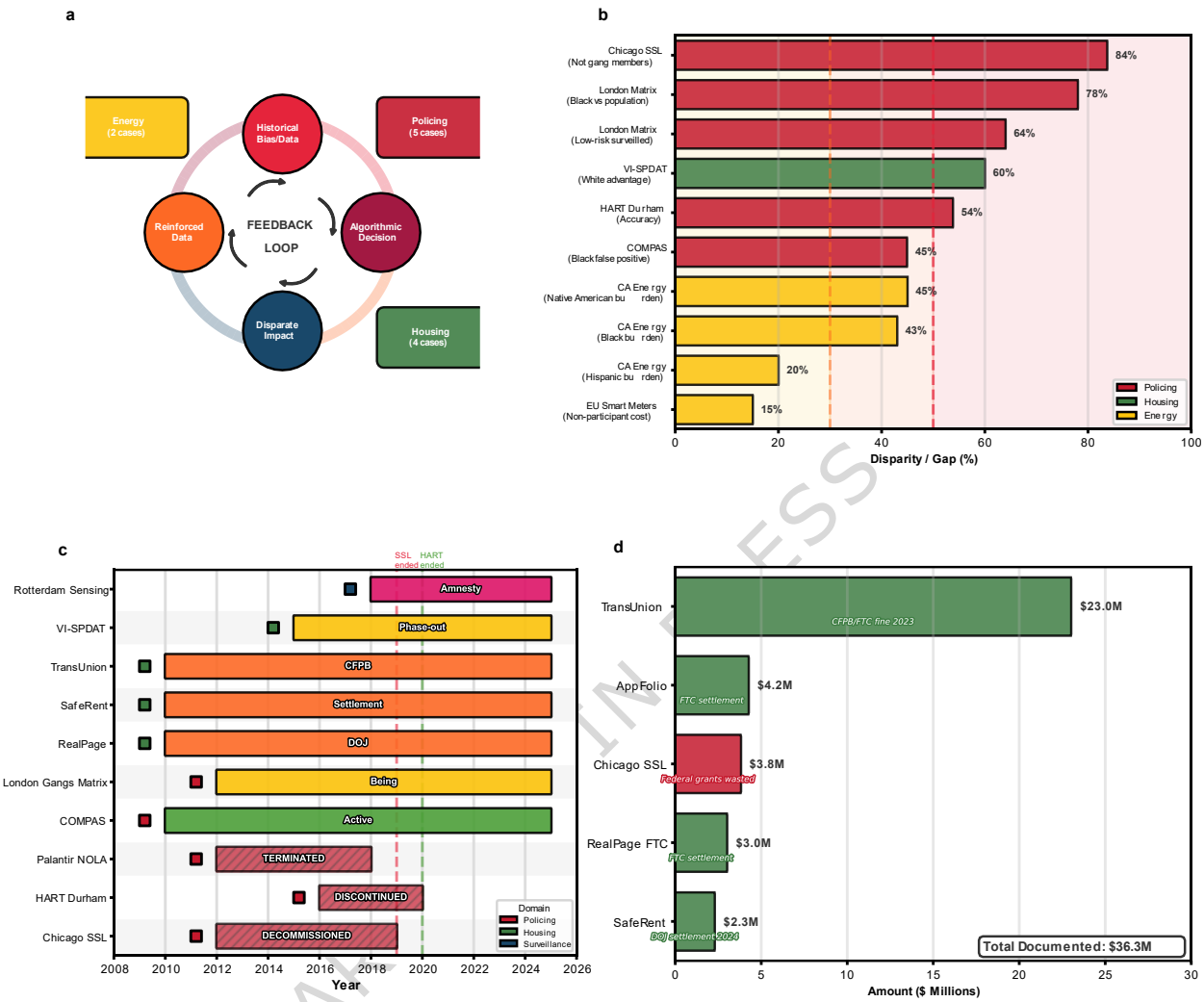

Figure 3 | Discriminatory Feedback Loops: Evidence from 11 Documented Cases. **a,** The feedback loop mechanism showing how historical bias feeds algorithmic decisions, which generate disparate impacts that reinforce biased data, with case distribution across policing (5), housing (4), and energy (2) domains. **b,** Documented disparities from case sources, ranging from Chicago SSL's 84% gap between algorithmic classification and verified status to EU Smart Meters' 15% cost penalty for non-participants. **c,** System status timeline showing implementation periods and key events including terminations (Chicago SSL 2019, Palantir NOLA 2018, HART Durham 2020) and regulatory interventions. **d,** Regulatory response totaling $36.3 million in documented fines and settlements across tenant screening systems.

Chicago's Strategic Subject List (SSL), operational from 2012-2019, illustrates how predictive policing systems can create feedback dynamics through recursive data generation. The system assigned risk scores from 1-500 to nearly 400,000 individuals based on attributes including arrest records and social network connections [41]. The Chicago Office of Inspector General found that only 16.3% of individuals on the list were confirmed gang members, despite district commanders believing the figure was approximately 95%, an 84 percentage point gap between algorithmic classification and verified status (Figure 3b). This discrepancy suggests how algorithmic outputs may acquire institutional authority that diverges from ground truth. Ensign et al.[33] developed formal models demonstrating how predictive

policing can create "runaway feedback loops" where increased police presence in flagged areas generates more arrests, which become new data points reinforcing original predictions regardless of underlying crime rate changes. The SSL was decommissioned in November 2019 after $3.8 million in federal grants, without documented evidence of violence reduction [41]. Similarly, the London Metropolitan Police's Gangs Matrix assigned "harm scores" to approximately 3,800 individuals, with the Information Commissioner's Office finding that 78% were Black despite Black Londoners comprising 13% of the population, and 64% of those listed were classified as lowest risk yet remained subject to surveillance [58].

New Orleans' covert partnership with Palantir Technologies from 2012-2018 demonstrates how opacity in algorithmic systems may compound feedback effects by preventing external verification or contestation. The partnership, exposed by The Verge in February 2018, operated without city council knowledge or public oversight, assessing approximately 3,900 individuals through a risk assessment database [59]. Criminal defense attorneys reported never receiving Palantir analytical products in discovery materials, raising questions about defendants' ability to challenge algorithmic bases for targeting. The program was terminated in March 2018 following public exposure. Rotterdam's Sensing Project, documented in Amnesty International's 2022 report, illustrates a related pattern through automated ethnic profiling. The system monitored vehicles and assigned risk scores specifically targeting what authorities termed "mobile banditry" allegedly associated with Eastern European nationals [60]. Amnesty documented that the system created conditions where vehicles flagged as potentially connected to Eastern Europeans faced enhanced monitoring, increasing likelihood of detecting irregularities that then became data points supporting original risk assessments, a pattern the organization concluded violated rights to privacy, data protection, and non-discrimination.

The tenant screening industry demonstrates how algorithmic feedback loops may operate through housing markets. RealPage (acquired for $9.6 billion in 2020) and CoreLogic (acquired for $6 billion in 2021) dominate a market where industry reports indicate widespread landlord adoption of screening software [61,62]. Rosen, Garboden, and Cossyleon[44] documented how algorithmic scoring may perpetuate discrimination through variables that correlate with race, such as credit histories shaped by historical lending disparities or addresses in previously redlined neighborhoods. When these systems deny housing, applicants may face housing instability affecting credit scores and rental history, potentially making future denials more likely. Federal Trade Commission settlements totaling $7.25 million against RealPage and AppFolio addressed accuracy and disclosure concerns [63,64], while the Department of Justice's lawsuit against SafeRent alleged these tools affect millions of rental applications [65]. The VI-SPDAT homelessness assessment tool, used across 16+ U.S. communities, exhibited documented racial disparities: multiple studies found white individuals were approximately 60% more likely to receive high prioritization scores compared to Black individuals assessed at similar vulnerability levels (Figure 3b).

Energy systems in our sample exhibit related patterns operating through different mechanisms. California's proposed income-graduated fixed charge of $24.15 monthly would apply uniformly regardless of consumption, with analysis by the American Council for an Energy-Efficient Economy documenting existing disparities: low-income households spend 8.1% of income on energy versus 2.3% for higher-income households, with Black households spending 43% more, Hispanic households 20% more, and Native American households 45% more of their income on energy [66]. The European Union's smart meter rollout, reaching 54% of households by 2021 with €47 billion invested [67], creates conditions where households unable to engage with smart technology, due to lacking smartphones, internet access,

or technical literacy, may face up to 15% higher costs through inability to shift consumption to off-peak periods. These seven cases suggest three potential feedback mechanisms operating across domains: enforcement feedback, where differential observation generates differential data (Chicago SSL, London Gangs Matrix); surveillance cascades, where initial algorithmic judgments expand into multiple institutional domains (Palantir, Rotterdam Sensing); and economic exclusion loops, where algorithmic denials or burdens may compound over time (tenant screening, energy systems). As Akpinar, De-Arteaga, and Chouldechova[68] demonstrated, even attempts to train algorithms on alternative data sources may not eliminate bias when structural factors shape data generation processes themselves. Table 2 summarizes documented disparities and mechanisms across these cases; we note that while these patterns appear consistent within our sample, their generalizability to other contexts requires further investigation.

Table 2. Documented Feedback Loop Mechanisms in Selected Cases

System	Domain	Initial Disparity Source	Documented Disparity	Feedback Mechanism	Status
Chicago SSL	Policing	Arrest history weighting	84% gap: 16.3% actual vs 95% believed gang members	Enforcement feedback	Decommissioned 2019
London Gangs Matrix	Policing	78% Black (13% of population)	64% lowest-risk still surveilled	Surveillance cascade	Active (under reform)
Palantir New Orleans	Policing	Covert operation without oversight	3,900 individuals assessed secretly	Opacity-enabled feedback	Terminated 2018
Rotterdam Sensing	Policing	Ethnic profiling of Eastern Europeans	Documented rights violations	Cross-border discrimination	Active (Amnesty criticism)
Tenant Screening	Housing	Historical rental/credit data	$36.3M in regulatory fines/settlements	Economic exclusion	Active (regulatory action)
VI-SPDAT	Housing	Assessment tool design	60% white advantage in prioritization	Scoring disparity	Phase-out recommended
CA Fixed Charges / EU	Energy	Income-blind structures	15-45% disparity by income/race	Digital divide amplification	Proposed / Active

Smart Meters					

Note: Disparities calculated from documented sources as cited in text.

Theme 3: Community Resistance and Democratic Pushback

Community resistance represents a countervailing dynamic in several documented cases, where grassroots organizing challenged algorithmic system deployments. Our analysis identified 20 cases exhibiting resistance patterns, resulting in seven documented system terminations or major reforms, ten municipal facial recognition bans, and two state-level warrant requirements (Figure 4). These cases cluster in contexts with robust civil society infrastructure: the United States accounts for the majority of documented resistance victories, with a notable concentration in Massachusetts where Boston's 2020 ban (passed 13-0) preceded adoption in five additional municipalities (Figure 4B). Coalition sizes ranged from grassroots networks of affected residents to formal alliances of 50+ organizations (Figure 4E). This section examines seven cases in depth to illustrate conditions and tactics associated with successful resistance, while noting that Singapore's traffic AI deployment, where no documented resistance occurred, may illuminate contextual factors that enable or constrain democratic pushback.

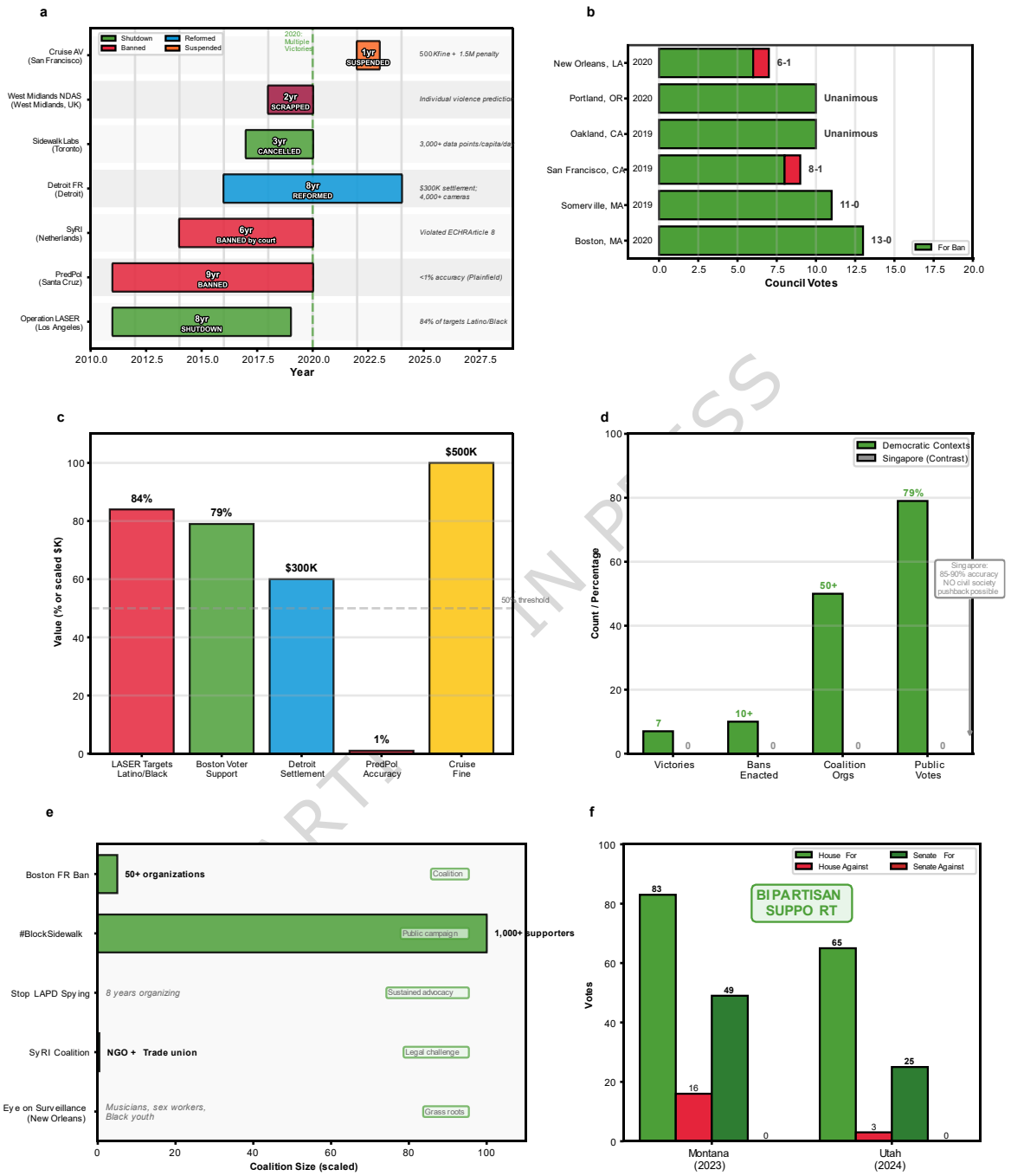

Figure 4 | Community Resistance and Democratic Pushback: Evidence from 20 Documented Cases. **a,** Timeline of resistance victories showing implementation periods, outcomes (shutdown/banned/reformed/suspended), and key documented harms. **b,** Facial recognition ban vote counts across U.S. municipalities, with Boston's 13-0 unanimous vote and subsequent regional diffusion to five Massachusetts cities. **c,** Key impact metrics including LASER's 84% Latino/Black targeting rate, Boston's 79% voter support, Detroit's $300,000 settlement, and Cruise's $500,000 federal fine. **d,** Democratic context comparison showing resistance indicators (victories, bans enacted, coalition

organizations, public support) in contexts where resistance occurred versus Singapore where no documented resistance emerged. **e,** Coalition composition and scale across five major campaigns, from Boston's 50+ organization coalition to grassroots movements in New Orleans. **f,** State-level warrant requirements for facial recognition in Montana (2023) and Utah (2024), both achieving bipartisan legislative majorities.

The defeat of Google's Sidewalk Labs smart city project in Toronto represents a case where sustained opposition resulted in complete project cancellation. The project, announced in 2017, proposed deploying sensor networks collecting data across Toronto's Quayside waterfront. Opposition led by the Canadian Civil Liberties Association (CCLA) and the #BlockSidewalk coalition, which mobilized over 1,000 supporters (Figure 4E), culminated in project cancellation in May 2020 [69]. The CCLA filed a constitutional challenge under the Charter of Rights and Freedoms, arguing the project raised privacy and democratic governance concerns [70]. The coalition achieved momentum when high-profile advisors resigned, including former Ontario Privacy Commissioner Ann Cavoukian, who departed citing privacy concerns. Similarly, the Netherlands' SyRI welfare fraud detection system, which cross-referenced personal data from multiple government databases and was deployed in low-income neighborhoods, was banned by court ruling in February 2020 after a coalition of NGOs and a trade union filed legal challenges arguing it violated European Convention on Human Rights Article 8 [60]. Both cases suggest that legal challenges combined with coalition organizing can achieve system termination, though we note these represent specific political and legal contexts that may not generalize.

Detroit's trajectory from facial recognition deployment to reform illustrates how individual cases of harm can catalyze policy change when connected to organized advocacy. The January 2020 wrongful arrest of Robert Williams, documented as the first U.S. case of facial recognition misidentification leading to arrest, became a focal point for organizing by Detroit Community Technology Project, ACLU Michigan, and allied organizations [45]. The June 2024 settlement achieved $300,000 compensation for Williams and established policies requiring corroborating evidence before arrests based on facial recognition matches (Figure 4C). Boston's facial recognition ban campaign, led by ACLU Massachusetts' "Press Pause Face Surveillance" initiative, achieved unanimous City Council passage (13-0) with documented support from 79% of Massachusetts voters in polling [46,71]. The campaign united over 50 organizations spanning racial justice, immigrant rights, and civil liberties groups (Figure 4E). Subsequently, Springfield, Cambridge, Northampton, Brookline, and Somerville adopted similar bans, creating regional policy diffusion within 18 months. At the state level, Montana (2023) and Utah (2024) enacted warrant requirements for police facial recognition use with bipartisan support, Montana's legislation passed 83-16 in the House and 49-0 in the Senate (Figure 4F).

The Los Angeles Police Department's 2019 termination of Operation LASER after eight years demonstrates how sustained community pressure can achieve system shutdown. The Stop LAPD Spying Coalition conducted community research documenting that 84% of individuals targeted by LASER were Latino or Black (Figure 4C), findings subsequently confirmed by the Inspector General's audit [72,73]. New Orleans' Palantir predictive policing partnership, operating covertly from 2012-2018 without city council knowledge, was terminated within months of The Verge's February 2018 exposé [59]. The revelation that criminal defense attorneys had never received Palantir analytical products in discovery materials intensified public pressure. These cases suggest that documentation of discriminatory impacts (LASER)

and exposure of secret operations (Palantir) can delegitimize systems, though the specific conditions enabling termination varied: LASER required eight years of sustained organizing, while Palantir collapsed rapidly once its covert nature was revealed.

Singapore's IBM traffic prediction system, deployed through the Land Transport Authority in 2006-2007 and achieving 85-90% accuracy [74], provides an instructive contrast that illuminates the relationship between system characteristics, democratic context, and resistance potential (Figure 4D). Critically, this system differs fundamentally from the person-based predictive systems documented in Themes 1-2: traffic flow optimization uses aggregate vehicle data to manage congestion rather than generating individualized risk scores that mark specific people for differential treatment [34,75]. This distinction matters because person-based prediction creates discriminatory feedback loops through recursive data generation, a dynamic largely absent in aggregate traffic optimization. We therefore do not suggest Singapore's traffic system warranted resistance comparable to systems like Chicago's SSL or COMPAS; the absence of documented opposition may appropriately reflect lower stakes for civil liberties.

However, Singapore's broader "Smart Nation" initiative does raise privacy concerns that have been documented by human rights organizations. Human Rights Watch [76] and CIVICUS[77] have characterized Singapore's civic space as "repressed," noting extensive surveillance networks, restrictions on public assembly under the Public Order Act, and laws that limit criticism of government systems. The smart city infrastructure promises efficiency, but as NYU's Center for Human Rights and Global Justice [78] documented, "the constant technology-driven surveillance and the loss of a few civil liberties are viewed by many as a small price to pay for such efficiency", a framing that forecloses rather than enables democratic deliberation about algorithmic governance. The comparison suggests that democratic infrastructure, civil society organizations, protected assembly rights, independent media, may be preconditions for the resistance patterns documented elsewhere, while acknowledging that the appropriateness of resistance depends on the specific harms a system produces. We caution against overgeneralizing from a single contrasting case. Table 3 summarizes implementation contexts and outcomes across these six cases; the patterns observed may reflect selection effects in our sample rather than universal dynamics of algorithmic resistance.

Table 3. Implementation Context and Resistance Outcomes in Selected Cases

Case	Period	Resistance Type	Key Actors	Outcome	Timeframe to Outcome
Toronto Sidewalk Labs	2017-2020	Coalition + legal challenge	CCLA, #BlockSidewalk (1,000+ supporters)	Complete cancellation	3 years
SyRI Netherlands	2014-2020	NGO coalition + legal challenge	NJCM, FNV Trade Union, Privacy First, civil society coalition	Court ban (ECHR Article 8 violation)	6 years
Detroit Facial Recognition	2016-2024	Individual case → systemic reform	ACLU Michigan, Detroit Community Tech	$300K settlement + strongest US FR policy	4 years
New Orleans Palantir*	2012-2018	Investigative exposure	The Verge, ACLU Louisiana	Program terminated	Months after exposure

Boston FR Ban	2019-2020	Proactive coalition	50+ organizations, 79% voter support	Ban passed 13-0, regional spread	1 year
LA Operation LASER	2011-2019	Research + sustained pressure	Stop LAPD Spying Coalition	Shutdown	8 years
Singapore Traffic AI	2006-present	No documented resistance	N/A	Continued operation	N/A (contrast case)

*Note: Outcomes and timeframes reflect documented sources; absence of documented resistance does not indicate public support. *: This case is common in Theme 2.*

Theme 4: Transparent Alternatives and Democratic AI Governance

Figure 5 maps transparency characteristics across urban AI implementations in our sample, revealing variation along multiple dimensions: process transparency (Figure 5a), documented performance outcomes (Figure 5b), the relationship between model opacity and governance openness (Figure 5c). Importantly, these cases occupy different positions on the transparency spectrum defined in our Introduction: some emphasize process transparency through documentation and registers (e.g., Amsterdam, Helsinki), others demonstrate contestability through stakeholder feedback mechanisms (e.g., Austin Energy's customer-facing interfaces, Seoul's waste collector feedback integration). The NYC Health + Hospitals case uniquely combines documented bias assessment with open-source methodology enabling independent verification. We present reported outcomes while acknowledging that cross-case comparisons are complicated by differences in urban context, implementation timeline, evaluation methodology, and data availability. Rather than claiming these cases prove transparent systems universally outperform opaque alternatives, a claim that would require controlled comparisons unavailable in real-world municipal deployments, we document the specific mechanisms through which transparency was operationalized and the outcomes that implementing organizations have reported.

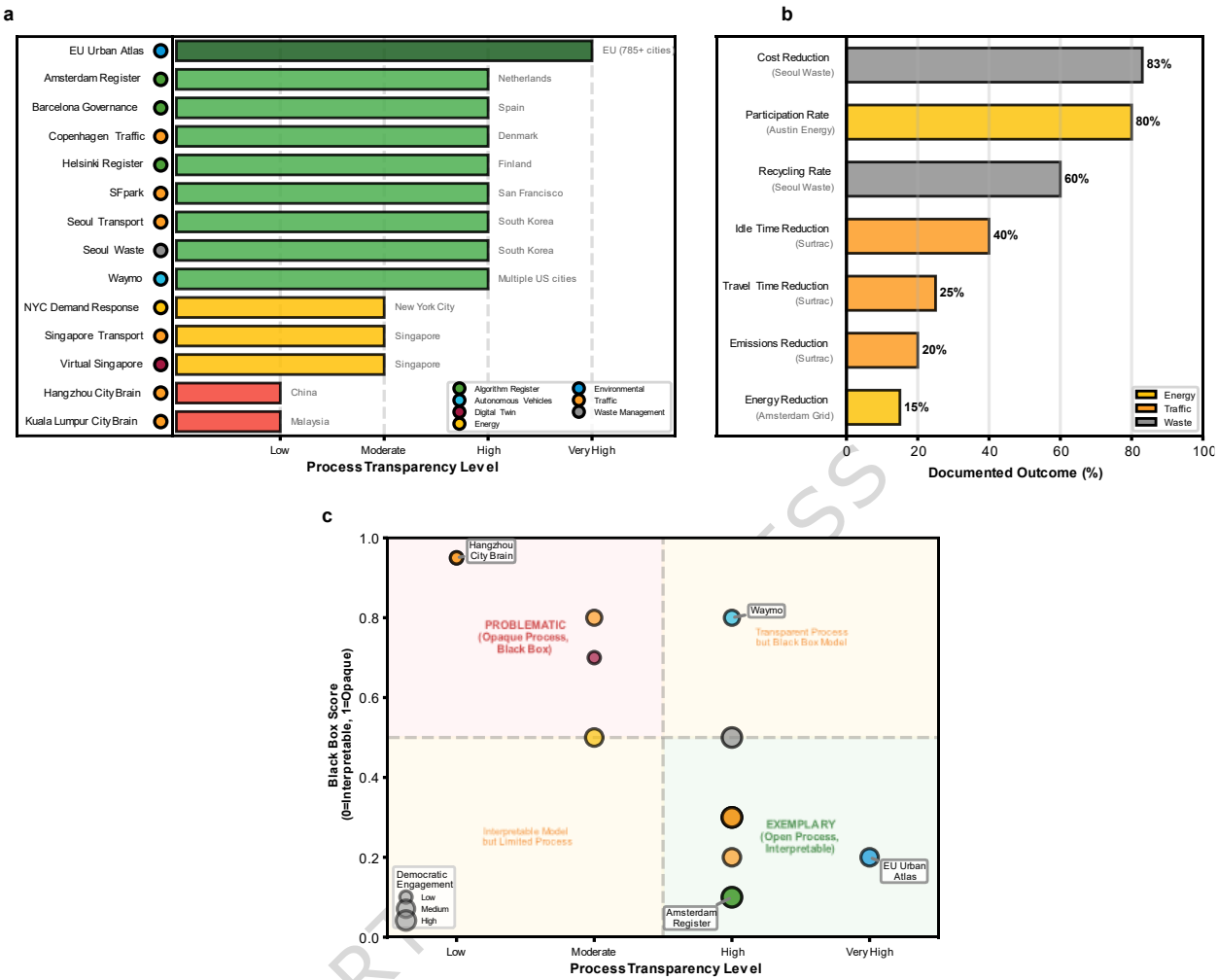

Figure 5 | The Transparency Spectrum in Urban AI Implementations. **a,** Process transparency levels across 16 implementations, ranked from low (Hangzhou City Brain) to very high (EU Urban Atlas, 785+ cities). **b,** Documented performance outcomes by domain: cost reduction (Seoul waste, 83%), participation rates (Austin Energy, 80%), recycling rates (Seoul, 60%+), and traffic improvements (Surtrac, 25-40%). **c,** Process transparency versus model opacity, with bubble size indicating democratic engagement level. Lower-right quadrant represents exemplary cases combining open governance with interpretable models.

Amsterdam and Helsinki launched municipal algorithm registers in September 2020, creating publicly accessible documentation of algorithmic systems used in city operations [47,79]. Amsterdam's register, implemented through the Saidot platform, documents systems from parking compliance to citizen report categorization, providing standardized fields for purpose, data sources, operational logic, and human oversight mechanisms [80]. Each entry includes plain-language descriptions intended to make algorithmic operations comprehensible to non-technical audiences. Amsterdam officials reported that the registration process prompted departments to articulate system purposes more explicitly, leading to identification of some redundancies, though no independent evaluation has quantified governance improvements [81]. Helsinki's parallel register documents rule-based chatbots and AI systems with feedback channels for citizen input [79]. Barcelona's Algorithmic Implementation Protocol, established in

2021, categorizes AI systems by risk level and mandates external auditing for high-risk applications [82]. The city's vCity project (2023-2025) integrates digital twin technology with the Decidim participatory platform, which has been adopted by over 200 cities globally, enabling residents to visualize algorithmic applications and propose modifications [83]. The EU's Living-in.EU Model AI Contractual Clauses (2023-2024), developed through 40+ expert roundtables, provide standardized procurement language for transparency requirements [84]. These institutional mechanisms represent attempts to operationalize algorithmic accountability, though their effectiveness in preventing harms depends on implementation fidelity, a factor we cannot assess from available documentation.

Seoul's smart waste management system, implemented through Ecube Labs beginning in 2014, reported an 83% reduction in collection costs following installation of 85 Clean Cube sensors with neural network-based route optimization [48,85]. The IoT sensors transmit real-time fill-level data accessible to collection staff through the Clean City Networks platform, and route optimization prioritizes bins exceeding fill thresholds while minimizing travel distances. Waste collectors provided operational feedback identifying constraints the algorithm initially missed, narrow alleys, market day obstacles, seasonal variations, enabling iterative refinement. The broader pay-as-you-throw system involved over 320 public meetings since 2001, with citizens' councils including environmental specialists contributing to system design; Seoul reports recycling rates exceeding 60% [48]. Austin Energy's demand response program, implemented through AutoGrid's DROMS platform, achieved 177 MW of demand response capacity with documented participation rates of 80% in smart thermostat pilots, compared to industry-reported averages of 20-30% for similar programs [86,87]. The system provides customer-facing interfaces displaying real-time consumption, predicted peak events, and potential savings. Austin Energy established a Low Income Consumer Advisory Task Force and provides CAP discounts for income-qualified customers, though independent assessment of equity outcomes has not been published. These cases illustrate how transparent operational logic may facilitate stakeholder engagement, though the contribution of transparency specifically, versus other factors such as program design, incentive structures, or local context, cannot be isolated from available data.

NYC Health + Hospitals, the largest municipal safety-net system in the United States serving over one million patients (90% patients of color, 70% Medicaid or uninsured), provides a documented example of proactive algorithmic bias assessment and mitigation in healthcare AI [49]. The system evaluated two binary classification models in their electronic medical record, one predicting acute visits for asthma and one predicting unplanned readmissions, across race/ethnicity, sex, language, and insurance status using Equal Opportunity Difference (EOD), a metric comparing false negative rates across subgroups. For the asthma model, baseline analysis revealed bias across race/ethnicity, with false negative rates ranging from 51% (Black or African American patients) to 82.8% (White patients). For the readmission model, insurance status showed the highest within-class burden of bias, with false negative rates ranging from 38.2% (Medicare) to 79.7% (Self-Pay). The system tested threshold adjustment, a post-processing method that sets subgroup-specific classification thresholds to minimize fairness metric differences, and successfully reduced absolute EOD to less than 5 percentage points for all targeted subgroups while maintaining accuracy losses below 10% and alert rate changes below 20%. The methodology, including open-source R code, was published with a supplementary playbook explicitly designed for other low-resource systems to replicate. This case demonstrates that bias identification and mitigation can be operationalized within resource-constrained settings, though the approach addresses post-hoc fairness correction rather than eliminating bias sources in training data or model architecture.

The European Space Agency's Urban Atlas processes 785-800 Functional Urban Areas across Europe using deep learning on Sentinel satellite imagery, achieving land use classification at continental scale [88]. Unlike many urban AI applications, the Urban Atlas operates under Copernicus open data policy (Regulation EU 1159/2013): all imagery, processing methods, and classification outputs are freely accessible, enabling verification, replication, and improvement proposals from any researcher or government. While deep learning classification involves technically opaque neural networks, the outputs are verifiable land cover categories that users can examine against source imagery, providing what might be termed "outcome transparency" even where "process transparency" of model internals is limited. Recent technical scholarship has examined relationships between model interpretability and predictive performance across specific application domains. Rudin[20] argues in Nature Machine Intelligence that the supposed accuracy- interpretability tradeoff is frequently overstated, providing evidence from domains including criminal justice risk assessment, medical diagnosis, and credit scoring where interpretable models matched black-box performance. In separate work, Rudin[89] and colleagues' analysis of NYC's electrical grid failure prediction, developed in collaboration with Con Edison over 2007-2012, demonstrated that machine learning models with interpretable features enabled utility engineers to identify failure modes and prioritize maintenance, with the transparent use of data allowing domain experts to troubleshoot and extend the models. Wagner et al.[27] demonstrated that gradient boosting decision tree models with SHAP values could capture up to 84% of variance in Berlin's trip emissions analysis while revealing policy-relevant threshold effects, such as the sharp increase in vehicle kilometers traveled beyond 10km from city centers, that inform targeted urban planning interventions. Kim et al.[28] showed that XGBoost- SHAP approaches in Seoul's urban expansion modeling achieved prediction accuracy exceeding 90% while providing actionable insights about land-cover characteristics driving growth patterns. These domain-specific findings suggest that transparency and performance need not be mutually exclusive in particular applications, though we note these represent specific contexts; whether the relationship generalizes across all urban AI deployment scenarios remains an empirical question requiring further investigation. The cases documented in this theme illustrate varied approaches to operationalizing transparency, from algorithm registers to bias mitigation playbooks to open data infrastructure, without claiming that any single approach constitutes a universal solution to algorithmic accountability challenges.

Discussion

The cross-case analysis reveals a recurring tension in urban AI implementations: technical performance metrics frequently diverge from policy outcomes, and this divergence appears structural rather than incidental. Across 157 documented cases, we observe patterns suggesting that the core challenge of urban AI governance is not primarily technical but epistemological, a systematic misalignment between what algorithms optimize for and what cities need them to achieve. This synthesis integrates findings across four themes to articulate principles for democratic urban AI governance, while acknowledging that these principles derive from our sample and require validation across broader contexts [29-31,90,91].

The accuracy illusion documented across our cases reveals what we term the "metrics trap": algorithmic systems achieving high performance on technical proxies while failing their stated policy objectives. The 88 percentage point gap between ShotSpotter's acoustic accuracy and crime-fighting effectiveness, COMPAS's aggregate accuracy masking disparate error rates, and healthcare algorithms optimizing for

cost while under-identifying patients for care demonstrate this common mechanism [41-43]. This pattern extends Muller [31]'s analysis of "the tyranny of metrics" into algorithmic governance, challenging techno-optimistic narratives [3,92] while providing explanatory mechanisms absent from purely critical accounts [10,93]. Breaking free requires redefining success through democratic deliberation, ensuring algorithms optimize for community-articulated goals rather than vendor-defined metrics [94-96].

The discriminatory feedback loops documented across domains reveal how algorithms can transform historical bias into "computational destiny." Chicago's Strategic Subject List exemplifies this: only 16.3% of listed individuals were confirmed gang members, yet district commanders believed 95%, demonstrating how algorithmic outputs acquire institutional authority independent of ground truth [33,41]. Our cases extend Lum and Isaac [34]'s analysis beyond predictive policing to housing, surveillance, and energy systems, suggesting feedback loops may represent a general property of algorithmic governance in unequal societies [6-8,44]. Interrupting these loops requires structural approaches: reforming data generation at source, incorporating affected community perspectives at decision points, and implementing regular audits with sunset provisions [97-99].

The community resistance cases reveal both the power and limitations of democratic opposition. Toronto, Boston, Detroit, and LA demonstrate that organized communities can achieve significant policy changes, but these campaigns required enormous resources and years of organizing [45,46,72]. The challenge lies in transforming episodic resistance into systematic governance, moving from what Hirschman [100] termed "voice" through protest toward institutionalized voice embedded in governance structures. Boston's ban diffusing to five Massachusetts municipalities suggests policy frameworks can scale through networked advocacy [38,39,101], while Singapore's contrast case illustrates how political context shapes these dynamics [37,102].

The transparent alternatives examined suggest that transparency can be operationalized through varied institutional mechanisms. Three distinct levels prove necessary for democratic accountability: technical transparency requiring interpretable models whose logic can be examined, moving beyond what Ananny and Crawford [21] term "seeing without knowing"; process transparency demanding visibility into development, procurement, and deployment [15]; and impact transparency necessitating ongoing monitoring of outcomes disaggregated by affected populations [25]. NYC Health + Hospitals' documented bias mitigation demonstrates these principles can be operationalized even in resource-constrained settings [49], while technical scholarship provides domain-specific evidence that interpretable models can achieve competitive performance [27,28,103].

Cities can institutionalize these insights through permanent structures combining bottom-up resistance with top-down policy frameworks. Barcelona's Advisory Council creates ongoing channels for community input [82], operationalizing Fung and Wright [36]'s "empowered participatory governance." The EU's Model AI Contractual Clauses demonstrate how procurement power can mandate transparency requirements [84], aligning with Ostrom [35]'s polycentric governance theory. The interaction between grassroots organizing and formal governance observed in Detroit suggests neither approach alone achieves sustainable accountability [104]. As Fung et al. [24] distinguish, transparency-as-disclosure differs from transparency-as-accountability; the latter requires mechanisms ensuring disclosed information generates consequences [47,80].

Several boundary conditions shape interpretation. Language constraints restricted analysis primarily to English-language sources. Publication bias likely favors documented failures over quietly successful

systems. Temporal constraints mean recent implementations lack comprehensive outcome data. Our critical orientation, while justified by documented harms, may undervalue genuine technical achievements; we view this stance as methodologically appropriate given power asymmetries while acknowledging its influence [15,32]. The cases examined represent specific political and institutional contexts, resistance succeeded predominantly where civil society infrastructure existed; transparent governance emerged in municipalities with prior participatory commitments, constraining direct generalization.

The patterns documented converge on a central finding: black-box optimization evaluated by technical metrics divorced from democratic accountability produces predictable harms across domains. Yet the transparent alternatives demonstrate this outcome is not inevitable. The evidence suggests cities need not accept supposed tradeoffs between technical capability and democratic governance. The path forward requires reconceptualizing algorithms not as neutral tools but as political systems requiring democratic oversight, extending Winner [16]'s insight that artifacts have politics into the algorithmic domain. Just as cities govern infrastructure and services through democratic processes, algorithmic systems must become subject to public accountability, community input, and ongoing assessment. The cases examined suggest this governance is not only necessary but achievable, though specific mechanisms require continued investigation across diverse contexts.

Methods

This study employs a multiple case study design integrated with thematic synthesis to examine the deployment and impacts of artificial intelligence systems across urban domains. This methodological approach was selected for its capacity to provide rich, contextual understanding of complex sociotechnical phenomena while enabling systematic cross-case analysis of patterns and themes [105,106].

Research Design Rationale

The multiple case study methodology is particularly suited to examining urban AI implementations for several reasons. First, AI deployments in cities represent bounded systems where technology, policy, and social outcomes intersect in specific contexts [107]. Second, the approach accommodates multiple data sources and types of evidence, reflecting the reality that urban AI systems are documented across academic, governmental, commercial, and journalistic sources [108]. Third, case study methodology explicitly supports critical examination of power relations and social justice concerns central to this investigation [109].

The integration of thematic synthesis enhances the analytical power of the case study approach by enabling systematic identification of patterns across cases while preserving contextual richness [110]. This combined methodology allows us to move beyond individual implementation stories to identify recurring mechanisms of success, failure, and resistance in urban AI deployment. The approach aligns with recent calls for methodological innovation in studying algorithmic systems, which require techniques capable of capturing both technical specifications and lived experiences of affected communities [111].

Researcher Positionality

As researchers trained in urban studies, data science, and human geography, we acknowledge our positionality shapes this investigation. We have technical expertise in machine learning systems combined with critical social science perspectives on urban inequality and digital justice. This interdisciplinary background provides sensitivity to both technical claims and social impacts, though we recognize it also predisposes us toward skepticism of technosolutionist narratives. We have previously engaged in community organizing around algorithmic accountability, which informs our commitment to centering affected communities' experiences. Throughout the research process, we maintained reflexive journals documenting how our perspectives evolved through engagement with the data, particularly when encountering cases that challenged our initial assumptions about the inevitability of algorithmic harm.

Case Universe and Sampling Strategy

Our case universe comprises 157 documented AI implementations across six urban domains spanning 27 countries from January 2015 to May 2024. This comprehensive mapping emerged from systematic searches of academic databases including Web of Science, Scopus, IEEE Xplore, ScienceDirect, and Google Scholar, supplemented by government repositories and verified news sources. The temporal boundaries capture the period following the deep learning revolution while providing sufficient time for implementation outcomes to manifest. We acknowledge this dataset reflects documented and accessible cases rather than a comprehensive global census; the concentration of cases in English-speaking and European contexts likely reflects language and documentation accessibility limitations in our search strategy rather than actual global deployment patterns.

From this universe, we selected 28 cases for in-depth thematic analysis through purposeful sampling [112]. Selection criteria prioritized documentation richness, defined as cases with evidence from at least four different source types including technical specifications, implementation reports, and community responses. We sought maximum variation across geographies, ensuring representation from Global North and Global South contexts, as well as variation in implementation outcomes from documented successes to documented failures. Following [107]'s critical case logic, we deliberately included extreme and paradigmatic cases that illuminate the boundaries of urban AI's possibilities and failures. Temporal coverage spanned early implementations from 2015-2018 through recent deployments in 2022-2024, enabling analysis of evolutionary patterns in the field.

Cases were included if they met all the following criteria:

The system must use machine learning, statistical models, or rule-based algorithms to process urban data and generate outputs affecting governance decisions or resident experiences. This includes:

- Predictive models (risk scoring, forecasting, classification)
- Optimization algorithms (routing, resource allocation, scheduling)
- Pattern recognition systems (image analysis, anomaly detection)
- Automated decision support systems
- Infrastructural platforms enabling multiple algorithmic applications

The system must be deployed (or seriously proposed for deployment) by municipal authorities, affect urban populations, or operate within city boundaries. Private systems (e.g., ride-sharing algorithms) were excluded unless directly integrated into public governance.

Sufficient documentation must exist across ≥4 source types to enable rigorous analysis. Sources must include technical specifications plus evidence of social impacts.

Implementation or serious deployment proposal must have occurred between January 2015 - May 2024 to capture post-deep-learning-revolution developments.

We deliberately included cases where proposed AI systems were successfully blocked or reformed through community resistance (Toronto Sidewalk Labs, Boston facial recognition ban, Detroit facial recognition reform, LA LASER termination, New Orleans Palantir termination). These cases are essential rather than peripheral because they:

- Reveal community capacity to identify algorithmic threats before or during implementation,
- Document effective resistance strategies applicable to future proposals,
- Show how democratic governance can prevent or reform harmful AI,
- Provide evidence against technological determinism.

Excluding resistance cases would systematically bias analysis toward systems that successfully overcame opposition, obscuring the democratic agency that shapes urban AI landscapes. Cases were excluded if they:

- Lacked sufficient documentation for rigorous analysis,
- Involved purely private-sector algorithms without public governance implications,
- Occurred primarily outside urban contexts (e.g., agricultural AI, wilderness monitoring),
- Were announced but never reached serious implementation stage (vaporware).

The final sample includes cases from multiple regions, with deliberate attention to implementations affecting marginalized communities whose experiences are often underdocumented in technical literature. This sampling strategy balances depth of analysis with breadth of coverage, exceeding [113]'s recommended range of 4-10 cases for theory building while remaining manageable for detailed qualitative analysis.

Data Collection and Management

Following Yin [106]'s principle of multiple evidence sources, we developed a comprehensive data collection protocol. Data sources encompassed academic publications including peer-reviewed articles, conference proceedings, and technical reports; government documents such as official reports, audits, policy documents, and regulatory findings; technical documentation from system specifications and vendor materials; news media, particularly investigative journalism documenting system failures and community impacts; legal documents including court filings, settlements, and regulatory enforcement actions; and community sources such as advocacy reports, public testimony, and documented resistance efforts.

This diverse source strategy reflects Stake [105]'s argument that case studies should capture multiple realities and perspectives. The inclusion of non-academic sources proves essential for documenting real-world impacts often absent from technical publications, particularly regarding marginalized communities' experiences [114]. For each case, we constructed a digital repository using Mendeley for academic sources and a structured filing system for grey literature, with all documents tagged by case, source type, and temporal period.

Data collection proceeded iteratively, beginning with technical documentation to understand system specifications, followed by government and academic sources for implementation details, and culminating with community sources and investigative journalism to capture impacts and resistance. This sequencing allowed us to identify gaps between technical claims and lived experiences.

Analytical Framework

Our analysis proceeded through three integrated phases, each building upon the previous to develop comprehensive understanding of urban AI implementations. The analytical process combined deductive frameworks from critical algorithm studies with inductive theme development from empirical data.

Each case underwent systematic analysis (phase 1) using NVivo 12 software. We developed an initial coding framework based on sociotechnical systems theory [14] and critical algorithm studies [6,8], encompassing technical architecture, stakeholder networks, implementation processes, documented outcomes, and resistance patterns. Two researchers independently coded five initial cases, achieving inter-rater reliability of 0.84 (Cohen's kappa) before refining the codebook. The refined framework was then applied to all cases, with regular team meetings to discuss emerging patterns and resolve interpretive differences.

Following Thomas and Harden [110]'s approach, we conducted line-by-line coding of case findings to develop descriptive themes (phase 2) staying close to primary data. Through constant comparison, these descriptive themes evolved into analytical themes capturing deeper patterns. We employed Braun & Clarke [115]'s reflexive thematic analysis principles, recognizing themes as analytical outputs rather than discovered entities. The synthesis generated 47 descriptive codes, consolidated into 12 analytical themes, and ultimately organized into four theoretical constructs: the accuracy illusion (Theme 1), discriminatory feedback loops (Theme 2), community resistance (Theme 3), and transparent alternatives (Theme 4).

Systematic cross-case analysis employed multiple techniques including pattern matching [106], explanation building [116], and qualitative comparative analysis principles [117]. We constructed detailed case matrices documenting implementation characteristics, outcomes, and contextual factors (phase 3). Pattern identification focused on mechanisms driving the "metrics trap" phenomenon, conditions enabling transitions to transparent systems, community resistance patterns and outcomes, and relationships between technical architectures and social impacts.

Quality and Rigor

We implemented multiple strategies to ensure analytical rigor and trustworthiness. Triangulation operated at multiple levels, comparing findings across data sources within cases, across cases within domains, and across domains for meta-patterns [118]. We maintained comprehensive audit trails in NVivo

documenting all coding decisions, theme development, and analytical memos totaling over 200 pages. This documentation enables traceability from raw data to theoretical claims.

Negative case analysis played a crucial role, with deliberate attention to implementations that defied emerging patterns. For instance, we extensively analyzed cases where black-box systems appeared to benefit communities and where transparent systems failed to prevent harm. These analyses refined our theoretical claims and identified boundary conditions. Member checking occurred through two mechanisms: sharing case summaries with documented implementation stakeholders (achieving responses from 31 of 67 contacted) and presenting preliminary findings at three practitioner conferences for feedback. Peer debriefing involved monthly meetings with an external advisory board of scholars in critical data studies, urban governance, and AI ethics who reviewed our analytical process and challenged interpretations.

Thematic Case Distribution

The 28 in-depth cases were organized across four analytical themes, with some cases appearing in multiple themes where their characteristics warranted examination from different analytical perspectives:

Theme 1 - The Accuracy Illusion (7 in-depth cases) includes: ShotSpotter Chicago and ShotSpotter NYC exemplify divergence between acoustic detection accuracy (97%) and policy effectiveness (9.1% of alerts leading to gun crime evidence in Chicago; 87% false alarm rate in NYC). COMPAS demonstrates how aggregate accuracy masks racially disparate error rates. The Obermeyer healthcare algorithm illustrates optimization for cost prediction while under-identifying Black patients for care. Delhi's air quality prediction and Boston University's hospitalization model show technical accuracy that may not capture differential impacts across populations. COVID-19 contact tracing applications, including Singapore's TraceTogether, achieved high proximity detection accuracy (70-95%) in laboratory conditions while failing to reach adoption thresholds necessary for population-level epidemic control."

Theme 2 - Discriminatory Feedback Loops (7 in-depth cases) contains: Chicago's Strategic Subject List (2012-2019) exemplifies algorithmic bias in predictive policing, with documented gaps between algorithmic classification and verified status. London's Gangs Matrix demonstrates regulatory intervention following documented racial disproportionality. New Orleans' Palantir partnership (2012-2018) reveals opacity in surveillance collaborations. Rotterdam's Sensing Project illustrates automated ethnic profiling. Tenant screening systems (RealPage, CoreLogic, SafeRent) demonstrate feedback dynamics in housing. VI-SPDAT homelessness assessment shows documented racial disparities in prioritization. Energy systems (California fixed charges, EU smart meters) illustrate digital divide amplification.

Theme 3 - Community Resistance (7 in-depth cases including one contrast case) includes: Toronto Sidewalk Labs cancellation represents successful coalition resistance against corporate smart city initiatives. Boston's facial recognition ban demonstrates preemptive policy action with regional diffusion. Detroit's facial recognition case shows transformation from documented harm to systemic reform. LA's LASER termination reveals how sustained community research and pressure achieved shutdown. New Orleans' Palantir termination demonstrates response to exposed covert operations. Singapore's traffic AI serves as contrast case where no documented resistance occurred, illustrating contextual factors that may enable or constrain democratic pushback.

Theme 4 - Transparent Alternatives (7 in-depth cases) includes: Amsterdam's algorithm register and Helsinki's AI register represent pioneering transparency infrastructure. Barcelona's AI governance framework including vCity and Decidim integration provides model for participatory governance. Seoul's waste management demonstrates IoT integration with worker feedback mechanisms. Austin Energy's demand response shows transparent customer-facing interfaces with documented participation rates. NYC Health + Hospitals' bias assessment and mitigation demonstrates operationalized algorithmic fairness in resource-constrained settings. EU Urban Atlas represents continental-scale open data infrastructure.

Several cases merit particular attention as critical or paradigmatic examples. Toronto's Sidewalk Labs cancellation demonstrates the power of sustained coalition resistance combining legal challenges, public advocacy, and expert resignations. Detroit's facial recognition case stands as paradigmatic for several reasons: it produced the first documented U.S. wrongful arrest from facial recognition, catalyzed unprecedented policy reform including mandatory corroborating evidence requirements, and achieved both individual remedy ($300,000 settlement) and systemic change. NYC Health + Hospitals' bias mitigation case demonstrates that algorithmic fairness assessment can be operationalized even in safety-net systems with limited resources, with open-source methodology enabling replication by other institutions.

Several notable implementations were excluded due to insufficient documentation or language barriers. Chinese social credit systems, while influential, lacked sufficient English-language documentation for rigorous analysis. Some European implementations in non-English speaking countries were similarly excluded despite potentially innovative approaches. Recent deployments (late 2024) lacked sufficient outcome data for meaningful evaluation of social impacts beyond technical claims.

This selection of 28 cases provides coverage across domains, geographies, and outcomes while maintaining analytical depth. The cases collectively demonstrate the patterns of metrics divergence, feedback dynamics, community resistance strategies, and transparency mechanisms that characterize contemporary urban AI deployment within our documented sample.

Data Availability

The raw data for this study is inside Data S1 and processed data for reproducing the findings of this study is available at *https://github.com/navid-nsk/metrics-trap*.

Code Availability

All the codes for analyzing the raw data for this study including generating all figures can be accessed at *https://github.com/navid-nsk/metrics-trap*.

Author Contributions

S.N.M.M. conceived and designed the study, developed the methodological framework, conducted the systematic case selection and data collection across 157 AI implementations, performed the qualitative analysis of 28 in-depth cases, created all figures and tables, wrote the manuscript, and approved the

final version for submission. H.C. supervised and participated in the research design, assessed and performed partially in qualitative analysis of 28 in-depth cases, participated in revision and writing the final manuscript version and approved the final version for submission.

Competing Interests

The authors declare no competing interests. None of the authors participated in any capacity as guest editor or reviewer to the journal or any collections.

Acknowledgements

We are grateful to Reviewer 1 for their insightful comments and appreciate Reviewer 2 for their constructive feedback. We would like to extend our special thanks to Reviewer 3 for their particularly valuable and thoughtful observation. No funding was received for this research.

References

1 Allam, Z. & Dhunny, Z. A. On big data, artificial intelligence and smart cities. *Cities* **89**, 80–91, doi:10.1016/j.cities.2019.01.032 (2019).

2 Yigitcanlar, T. *et al.* Understanding 'smart cities': Intertwining development drivers with desired outcomes in a multidimensional framework. *Cities* **81**, 145–160, doi:10.1016/j.cities.2018.04.003 (2020).

3 Batty, M. *The new science of cities*. (MIT Press, 2013).

4 Yigitcanlar, T. *et al.* Unlocking artificial intelligence adoption in local governments: Best practice lessons from real-world implementations. *Smart Cities* **7**, 1576-1625 (2024).

5 Organisation for Economic Co-operation and Development. *Artificial intelligence for advancing smart cities (OECD Smart Cities and Inclusive Growth Issues Note)*, <https://www.oecd.org/content/dam/oecd/en/about/programmes/cfe/the-oecd-programme-on-smart-cities-and-inclusive-growth/Issues-Note-AI-for-advancing-smart-cities.pdf> (2025).

6 Benjamin, R. *Race after technology: Abolitionist tools for the new Jim code*. (Polity, 2019).

7 Eubanks, V. *Automating inequality: How high-tech tools profile, police, and punish the poor*. (Martin's Press, 2018).

8 Noble, S. U. *Algorithms of oppression: How search engines reinforce racism*. (NYU Press, 2018).

9 O'neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. (Crown, 2017).

10 Greenfield, A. *Against the Smart City: A Pamphlet. This is Part I of" The City is Here to Use"*. (Do projects, 2013).

11 Hollands, R. G. in *The Routledge companion to smart cities* 179-199 (Routledge, 2020).

12 Kitchin, R. *The data revolution: Big data, open data, data infrastructures and their consequences*. (SAGE, 2014).

13 Söderström, O., Paasche, T. & Klauser, F. in *The Routledge companion to smart cities* 283-300 (Routledge, 2020).

14 Bijker, W. E.,, Hughes, T. P., & Pinch, T. . *The social construction of technological systems: New directions in the sociology and history of technology*. Anniversary edn, (MIT Press, 2012).

15 Pasquale, F. in *The black box society* (Harvard university press, 2015).

16 Winner, L. in *Computer ethics* 177-192 (Routledge, 2017).
17 Lipton, Z. C. The mythos of model interpretability. *Queue* **16**, 31–57, doi:10.1145/3236386.3241340 (2018).
18 Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1798–1828, doi:10.1109/TPAMI.2013.50 (2013).
19 Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* **3**, 2053951715622512 (2016).
20 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215, doi:10.1038/s42256-019-0048-x (2019).
21 Ananny, M. & Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* **20**, 973-989 (2018).
22 Rudin, C. *et al.* Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* **16**, 1-85 (2022).
23 Molnar, C. *Interpretable machine learning: A guide for making black box models explainable*. (Lulu.com, 2022).
24 Fung, A., Graham, M. & Weil, D. *Full disclosure: The perils and promise of transparency*. (Cambridge University Press, 2007).
25 Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. & Vertesi, J. in *Proceedings of the conference on fairness, accountability, and transparency.* 59-68.
26 Kroll, J. A. *et al.* Accountable algorithms, 165 U. *Pa. L. Rev* **633**, 633 (2017).
27 Wagner, F. *et al.* Using explainable machine learning to understand how urban form shapes sustainable mobility. *Transportation Research Part D: Transport and Environment* **111**, 103442, doi:10.1016/j.trd.2022.103442 (2022).
28 Kim, M., Kim, D., Jin, D. & Kim, G. Application of explainable artificial intelligence (XAI) in urban growth modeling: A case study of Seoul Metropolitan Area, Korea. *Land* **12**, 420, doi:10.3390/land12020420 (2023).
29 Strathern, M. 'Improving ratings': audit in the British University system. *European review* **5**, 305-321 (1997).
30 Thomas, R. L. & Uminsky, D. Reliance on metrics is a fundamental challenge for AI. *Patterns* **3**, 100476, doi:10.1016/j.patter.2022.100476 (2022).
31 Muller, J. Z. The tyranny of metrics: The quest to quantify everything undermines higher education. *The Chronicle of Higher Education* **64**, 1-7 (2018).
32 Jasanoff, S. *States of knowledge*. (Taylor & Francis Abingdon, UK, 2004).
33 Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research* **81**, 160–171 (2018).
34 Lum, K. & Isaac, W. To predict and serve? *Significance* **13**, 14-19 (2016).
35 Ostrom, E. *Governing the commons: The evolution of institutions for collective action*. (Cambridge university press, 1990).
36 Fung, A. & Wright, E. O. Countervailing Power in Empowered Participatory. *Deepening democracy: Institutional innovations in empowered participatory governance* **4**, 259 (2003).
37 McCarthy, J. D. & Zald, M. N. Resource mobilization and social movements: A partial theory. *American journal of sociology* **82**, 1212-1241 (1977).
38 Dolowitz, D. P. & Marsh, D. Learning from abroad: The role of policy transfer in contemporary policy-making. *Governance* **13**, 5-23 (2000).

39 Walker, J. L. The diffusion of innovations among the American states. *American political science review* **63**, 880-899 (1969).

40 Snow, D. A., Rochford Jr, E. B., Worden, S. K. & Benford, R. D. Frame alignment processes, micromobilization, and movement participation. *American sociological review*, 464-481 (1986).

41 City of Chicago Office of Inspector General. *The Chicago Police Department's use of ShotSpotter technology*, <https://igchicago.org/2021/08/24/oig-finds-that-shotspotter-alerts-rarely-lead-to-evidence-of-a-gun-related-crime-and-that-presence-of-the-technology-changes-police-behavior/> (2021).

42 Angwin, J., Larson, J., Mattu, S. & Kirchner, L. *Machine bias: There's software used across the country to predict future criminals*, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (2016).

43 Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453, doi:10.1126/science.aax2342 (2019).

44 Rosen, E., Garboden, P. M. E. & Cossyleon, J. E. Racial discrimination in housing: How landlords use algorithms and home visits to screen tenants. *American Sociological Review* **86**, 787–822, doi:10.1177/00031224211029618 (2021).

45 A.C.L.U. *Civil rights advocates achieve the nation's strongest police department policy on facial recognition technology*. (American Civil Liberties Union, ACLU of Michigan, & University of Michigan Law School Civil Rights Litigation Initiative, 2024).

46 A.C.L.U. Massachusetts. *Press pause on face surveillance*. (American Civil Liberties Union of Massachusetts, 2020).

47 Amsterdam Municipality. *Amsterdam Algorithm Register [Beta version*. (City of Amsterdam, Chief Technology Office, 2020).

48 Seoul Metropolitan Government. Smart waste management system implementation results. *Seoul* (2014).

49 Mackin, S., Major, V. J., Chunara, R. & Newton-Dame, R. Identifying and mitigating algorithmic bias in the safety net. *npj Digital Medicine* **8**, 335 (2025).

50 Bronars, S. G. & Lynch, G. *Independent Audit of the ShotSpotter Accuracy, 2019-2022*, <https://www.edgeworthconomics.com/experience-independent-audit-of-the-shotspotter-accuracy> (2022).

51 Doucette, M. L., Green, C., Necci Dineen, J., Shapiro, D. & Raissian, K. M. Impact of ShotSpotter technology on firearm homicides and arrests among large metropolitan counties: A longitudinal analysis, 1999-2016. *Journal of Urban Health* **98**, 609–621, doi:10.1007/s11524-021-00515-4 (2021).

52 Piza, E. L. Gunshot detection technology time savings and spatial precision: An exploratory analysis in Kansas City. *Criminal Justice Policy Review* **34**, 3–23, doi:10.1177/08874034221110952 (2023).

53 Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* **5**, 153-163 (2017).

54 Kleinberg, J., Ludwig, J., Mullainathan, S. & Sunstein, C. R. Discrimination in the Age of Algorithms. *Journal of legal analysis* **10**, 113-174 (2018).

55 Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B. & Selvarajan, S. Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports* **14**, 6795, doi:10.1038/s41598-024-54807-1 (2024).

56 Hao, B. *et al.* Early prediction of level-of-care requirements in patients with COVID-19. *Elife* **9**, e60519 (2020).

57 Yi, F., Xie, Y. & Jamieson, K. in *Proceedings of the 2nd Workshop on Deep Learning for Wellbeing Applications Leveraging Mobile Devices and Edge Computing.* 1-6.

58 I.C.O. *Enforcement notice: Metropolitan Police Service*. (Information Commissioner's Office, 2018).

59 Winston, A. *Palantir has secretly been using New Orleans to test its predictive policing technology*. (The Verge, 2018).

60 Amnesty International. *We sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands*. (Amnesty International, 2022).

61 Bravo, T. *Thoma Bravo to acquire RealPage for $9.6 billion*. (Business Wire, 2020).

62 CoreLogic. *Stone Point Capital and Insight Partners to acquire CoreLogic for $6 billion*. (CoreLogic Press Release, 2021).

63 F.T.C. *Texas company will pay $3 million to settle FTC charges*. (Federal Trade Commission, 2018).

64 F.T.C. Tenant background report provider settles FTC allegations. *Federal Trade Commission*, doi:https://www.ftc.gov/news-events/news/press-releases/2020/12/ (2020).

65 Louis v. Saferent Sols. Vol. 685 F.Supp.3d 19 (U.S. District Court , District of Massachusetts, 2023).

66 Drehobl, A., Ross, L. & Ayala, R. *How high are household energy burdens? An assessment of national and metropolitan energy burden across the United States*. (American Council for an Energy-Efficient Economy, 2020).

67 ACER. Market Monitoring Report - Energy Retail and Consumer Protection. (European Union Agency for the Cooperation of Energy Regulators, 2021).

68 Akpinar, N. J., De-Arteaga, M. & Chouldechova, A.

69 CBC News. *Sidewalk Labs cancels plan to build high-tech neighbourhood in Toronto amid COVID-19*, <https://www.cbc.ca/news/canada/toronto/sidewalk-labs-cancels-project-1.5559370> (2020).

70 C.C.L.A. *CCLA v. Waterfront Toronto, et al*. (Canadian Civil Liberties Association, 2020).

71 Boston City Council. *Ordinance banning facial recognition technology in Boston*. (City of Boston, 2020).

72 Brennan Center for Justice. *Predictive policing explained*, <https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained> (2024).

73 CBS News. *LAPD still all-in on data-driven policing after scrapping controversial LASER program*, <https://www.cbsnews.com/news/los-angeles-police-department-laser-data-driven-policing-racial-profiling-2-0-cbsn-originals-documentary/> (2024).

74 IBM. *IBM and Singapore's Land Transport Authority pilot innovative traffic prediction tool*. (IBM Press Release, 2007).

75 Ferguson, A. G. in *The rise of big data policing* (New York University Press, 2017).

76 Human Rights Watch. *Singapore: Laws chill free speech, assembly*, <https://www.hrw.org/news/2017/12/13/singapore-laws-chill-free-speech-assembly> (2017).

77 CIVICUS. *People power under attack 2024 (CIVICUS Monitor Annual Report)*, <https://civicusmonitor.contentfiles.net/media/documents/GlobalFindings2024.EN.pdf> (2024).

78 Chandrasekhar, R. *Singapore's "smart city" initiative: One step further in the surveillance, regulation and disciplining of those at the margins*, <https://chrgj.org/2022-03-18-singapore-smart-city-initiative/> (2022).

79 City of Helsinki. *Helsinki AI Register*. (City of Helsinki, Digital and Population Data Services Agency, 2020).

80 Saidot. *Why AI governance is good for the bottom line*, <https://www.saidot.ai/insights/why-ai-governance-is-good-for-the-bottom-line> (2025).

81 Elliot. *Amsterdam Built the 'Perfect' Ethical AI System. It Still Failed. Here's Why.*, <https://medium.com/@elliotJL/amsterdam-built-the-perfect-ethical-ai-system-it-still-failed-here-s-why-8dc8072beea3> (2025).

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

83 OECD-OPSI. *vCity, a human-centric platform for urban digital twins*. (Observatory of Public Sector Innovation, 2024).

84 EU Public Buyers Community. *Updated EU AI model contractual clauses*, <https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/updated-eu-ai-model-contractual-clauses> (2024).

85 Ecube Labs. *Seoul Metropolitan Government case study*. (Ecube Labs, 2015).

86 Austin Energy.

87 AutoGrid Systems. *DROMS case study: Austin Energy*. (AutoGrid).

88 E.S.A./Copernicus. *Urban Atlas 2018 - Land cover/land use*. (European Space Agency/Copernicus Land Monitoring Service, 2021).

89 Rudin, C. *et al.* Machine learning for the New York City power grid. *IEEE transactions on pattern analysis and machine intelligence* **34**, 328-345 (2011).

90 Campbell, D. T. Assessing the impact of planned social change. *Evaluation and program planning* **2**, 67-90 (1979).

91 Porter, T. M. Trust in numbers: The pursuit of objectivity in science and public life. (2020).

92 Townsend, A. M. *Smart cities: Big data, civic hackers, and the quest for a new utopia*. (WW Norton & Company, 2013).

93 Morozov, E. *To save everything, click here: The folly of technological solutionism*. (PublicAffairs, 2013).

94 Fung, A. Empowered participation: Reinventing urban democracy. (2009).

95 Smith, G. *Democratic innovations: Designing institutions for citizen participation*. (Cambridge University Press, 2009).

96 Cardullo, P. & Kitchin, R. Being a 'citizen'in the smart city: Up and down the scaffold of smart citizen participation in Dublin, Ireland. *GeoJournal* **84**, 1-13 (2019).

97 Citron, D. K. & Pasquale, F. The scored society: Due process for automated predictions. *Wash. L. Rev.* **89**, 1 (2014).

98 Brayne, S. Big data surveillance: The case of policing. *American sociological review* **82**, 977-1008 (2017).

99 Collins, H. M. & Evans, R. The third wave of science studies: Studies of expertise and experience. *Social studies of science* **32**, 235-296 (2002).

100 Hirschman, A. O. Exit, Voice, and. *Loyalty: Responses to Decline in* (1970).

101 Keck, M. E. & Sikkink, K. A. *Activists beyond borders: Advocacy networks in international politics*. (Cornell University Press, 2014).

102 Tarrow, S. *Power in movement*. (Cambridge university press, 2022).

103 Rudin, C. & Radin, J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review* **1**, doi:10.1162/99608f92.5a8a3a3d (2019).

104 Levi-Faur, D. *The Oxford handbook of governance*. (Oxford University Press, 2012).

105 Stake, R. E. *Multiple case study analysis*. (The Guilford Press, 2006).

106 Yin, R. K. *Case study research and applications: Design and methods*. 6th edn, (SAGE Publications, 2018).

107 Flyvbjerg, B. Five misunderstandings about case-study research. *Qualitative Inquiry* **12**, 219–245, doi:10.1177/1077800405284363 (2006).

108 Bowen, G. A. Document analysis as a qualitative research method. *Qualitative Research Journal* **9**, 27–40, doi:10.3316/QRJ0902027 (2009).

109 Bartlett, L. & Vavrus, F. (2016).

110 Thomas, J. & Harden, A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology* **8, Article 45**, doi:10.1186/1471-2288-8-45 (2008).

111 Katzenbach, C. & Ulbricht, L. Algorithmic governance. *Internet Policy Review* **8**, 1–18, doi:10.14763/2019.4.1424 (2019).

112 Patton, M. Q. *Qualitative research and evaluation methods: Integrating theory and practice*. 4th edn, (SAGE Publications, 2015).

113 Eisenhardt, K. M. Building theories from case study research. *Academy of Management Review* **14**, 532–550, doi:10.5465/amr.1989.4308385 (1989).

114 Gillborn, D., Warmington, P. & Demack, S. QuantCrit: Education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity and Education* **21**, 158–179, doi:10.1080/13613324.2017.1377417 (2018).

115 Braun, V. & Clarke, V. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* **11**, 589–597, doi:10.1080/2159676X.2019.1628806 (2019).

116 Miles, M. B., Huberman, A. M. & Saldaña, J. *Qualitative data analysis: A methods sourcebook*. 4th edn, (SAGE Publications, 2020).

117 Ragin, C. C. *The comparative method: Moving beyond qualitative and quantitative strategies (With a new introduction*. (University of California Press, 2014).

118 Denzin, N. K., & Lincoln, Y. S. . *The SAGE handbook of qualitative research*. 5th edn, (SAGE Publications, 2017).